Quelques contributions aux méthodes numériques probabilistes et à la modélisation stochastique

Jérôme Lelong

Université Grenoble Alpes

Habilitation à diriger des recherches

présentée par

Jérôme Lelong

Spécialité: Informatique et Mathématiques appliquées

Quelques contributions aux méthodes numériques probabilistes et à la modélisation stochastique

Soutenue le 28 septembre 2017 devant le jury composé de

M. Anatoli JUDITSKY	**Examinateur**
Mme. Adeline LECLERCQ SAMSON	**Examinatrice**
M. Eric MOULINES	**Rapporteur**
M. Gilles PAGÈS	**Président**
M. John SCHOENMAKERS	**Rapporteur**
M. Denis TALAY	**Rapporteur**

A l'heure d'apporter une touche finale à ce manuscrit, je tiens à remercier Eric Moulines, John Schoenmakers et Denis Talay d'avoir accepté de rapporter mon habilitation. Je suis très honoré qu'ils aient pris le temps de lire ce document en détail en dépit d'un calendrier administratif parfois serré en période estivale. Merci également à Anatoli Juditsky, Gilles Pagès et Adeline Samson de participer à mon jury, je les en remercie chaleureusement.

La rédaction ce cette HDR est l'aboutissement de nombreuses années de travail passées au sein du Laboratoire Jean Kuntzmann, 8 ans déjà. Je tiens à remercier tous ses membres qui contribuent à en faire un cadre de travail agréable. Merci à Jérôme pour ses conseils avisés en cette période de préparation de la soutenance. Merci à Stéphane pour son aide à surmonter les différentes péripéties administratives. Puisque le travail n'est pas fait que de travail (chacun connaît l'importance de la machine à café), je me permets ici quelques *special thanks*. Merci à Fred, Edouard et Rémy pour nos nombreuses discussions de *geek*. Merci à Franck et Fred pour nos interminables discussions vélo. Je profite de cette occasion pour saluer les assidus de la cafétéria avec qui j'ai partagé tant de pauses méridiennes : Caroline, Delphine, Franck, Fred, Laurence, Stéphane, ...

Je voudrais également remercier tous les collègues avec qui j'ai eu la chance de collaborer que ce soit pour attaquer à un nouveau problème, monter un projet ou encore organiser un colloque : Agnès, Ahmed, Antonino, Aurélien, Benjamin, Bernard, Franck, Jean-Philippe, Jeff, Marianne, Stéphane et les nombreux autres que j'oublie forcément. Pardon à eux.

Je terminerais en saluant mes collègues de l'Ensimag et en particulier Sonia et Mnacho qui ont su me décharger de la gestion de la filière *Ingénierie pour la Finance* lorsque j'avais besoin de temps pour avancer la rédaction de ce manuscrit.

Pour conclure, très grand merci à Céline pour son soutien sans faille et tout particulièrement ces dernières semaines durant lesquelles mon attention quotidienne s'est considérablement réduite, la faute à une Haute Disponibilité Requise par ailleurs.

Contents

Publications

Only the papers marked with ♠ are presented in this document.
All my papers are available from my webpage
`http://www-ljk.imag.fr/membres/Jerome.Lelong/papers.html`.

Published papers

♠ [L-1] A. Kebaier and J. Lelong. Coupling Importance Sampling and Multilevel Monte Carlo using Sample Average Approximation. *Methodology and Computing in Applied Probability*, 2017, `http://dx.doi.org/10.1007/s11009-017-9579-y`.

[L-2] A. Alfonsi, C. Labart, and J. Lelong. Stochastic local intensity loss models with interacting particle systems. *Math. Finance*, 26(2):366–394, 2016 (Online 2013).

[L-3] C. De Luigi, J. Lelong, and S. Maire. Adaptive numerical integration and control variates for pricing basket options. *Applied Numerical Mathematics*, 100:14–30, 2016.

[L-4] M. Clausel, J.-F. Coeurjolly, and J. Lelong. Stein estimation of the intensity of a spatial homogeneous Poisson point process. *Ann. Appl. Probab.*, 26(3):1495–1534, 2016.

♠ [L-5] L. Badouraly Kassim, J. Lelong, and I. Loumrhari. Importance sampling for jump processes and applications to finance. *Journal of Computational Finance*, 19(2), 2015.

♠ [L-6] P. Etoré, S. Labbé, and J. Lelong. Long time behaviour of a stochastic nano particle. *Journal of Differential Equations*, 257(6):2115–2135, 2014.

♠ [L-7] J. Lelong. Asymptotic normality of randomly truncated stochastic algorithms. *ESAIM. Probability and Statistics*, 17:105–119, 2013.

[L-8] C. Labart and J. Lelong. A parallel algorithm for solving BSDEs. *Monte Carlo Methods Appl.*, 19(1), 2013.

[L-9] A. Alfonsi and J. Lelong. A closed form extension to the Black–Cox model. *Int. Journal of Theo. and Appl. Finance*, 15(8):30, 2012.

[L-10] J.-P. Chancelier, B. Lapeyre, and J. Lelong. Using premia and nsp for constructing a risk management benchmark for testing parallel architecture. *Concurrency and Computation: Practice and Experience*, 26(9), 2012.

♠ [L-11] B. Lapeyre and J. Lelong. A framework for adaptive Monte–Carlo procedures. *Monte Carlo Methods Appl.*, 17(1), 2011.

♠ [L-12] B. Jourdain and J. Lelong. Robust Adaptive Importance Sampling for Normal Random Vectors. *Ann. Appl. Probab.*, 19(5):1687–1718, 2009.

[L-13] C. Labart and J. Lelong. Pricing Double Parisian options using Laplace transforms. *Int. Journal of Theo. and Appl. Finance*, 12(1):19–44, 2009.

[L-14] C. Labart and J. Lelong. Pricing parisian options using laplace transforms. *Bankers, Markets Investors*, 99:29–43, March-April 2009.

♠ [L-15] J. Lelong. Almost sure convergence of randomly truncated stochastic algorithms under verifiable conditions. *Statistics & Probability Letters*, 78(16), 2008.

Book chapters, conference proceedings

[L-16] J. Lelong and A. Zanette. *Tree Methods*. John Wiley & Sons Ltd., 2010.

[L-17] J.-P. Chancelier, B. Lapeyre, and J. Lelong. Using premia and nsp for constructing a risk management benchmark for testing parallel architecture. In *Proceedings of the 2009 IEEE International Symposium on Parallel&Distributed Processing*, pages 1–6, Washington, DC, USA, 2009. IEEE Computer Society.

[L-18] J. Lelong. Truncated stochastic algorithm and variance reduction:toward an automatic procedure. In *Proceedings of the 6th International Workshop on Rare Event Simulation*, Bamberg, Germany, 2006.

Library

♠ [L-19] J. Lelong. *PNL : An open source scientific Library*. `https://pnlnum.github.io/pnl/`, 2007–2017.

Phd Thesis

[L-20] J. Lelong. *Asymptotic properties of stochastic algorithms and pricing of Parisian options*. PhD thesis, Ecole Nationale des Ponts et Chaussées, `https://tel.archives-ouvertes.fr/tel-00201373/`, September 2007.

Submitted papers

♠ [L-21] J. Lelong. Pricing American options using martingale bases. In revision to *SIAM J. Financial Mathematics*, `https://hal.archives-ouvertes.fr/hal-01299819`, 2016.

♠ [L-22] S. Labbé and J. Lelong. Stochastic modelling of thermal effects on a ferromagnetic nano particle. Technical report, `https://hal.archives-ouvertes.fr/hal-01337197`, 2016.

Chapter 1

Introduction

This document presents a survey of my research work since I defended my PhD in 2007. Even though I have been working a wide variety of subjects, my work basically falls in one of the following two categories.

- Either, my papers propose and study new numerical methods for stochastic problems. These papers are supported by a competitive C++ implementation of the proposed algorithms. When the size of the targeted problems justified it, a parallel algorithm and implementation were even proposed. See Sections 1.1 and 1.2.
- Or they target modeling issues and study how stochastic aspects can help modeling a given phenomenon. See Section 1.3. My works on modeling have mainly tackled two applications: finance and ferromagnetism. To keep this document coherent, I have chosen to only expose my results on stochastic modeling for ferromagnetism as the mathematical analysis of the model uses techniques related to stochastic approximation.

This document presents 9 articles of mine, which are all somehow related to stochastic optimization.

1.1 Adaptive numerical methods

Making numerical methods adaptive is usually achieved by solving a stochastic optimization problem to automatically learn the context.

Consider the problem of computing $\mathbb{E}[f(X)]$ by a Monte Carlo method and assume it admits a parametric representation of the form $\mathbb{E}[f(X)] = \mathbb{E}[h(\theta, Y)]$ for all θ. The key motivation for 6 articles of mine was to make the most of this free parameter, knowing that the goal is to choose the value θ^* minimizing $v(\theta) = \mathbb{E}[h(\theta, Y)^2]$. When this parametric representation comes from the use of importance sampling, we proved that the function v was of class C^1 and that we could find a function H such that $\nabla v(\theta) = \mathbb{E}[H(\theta, Y)]$.

Two methods can be used to approximate θ^* in such a framework: stochastic approximation (see [37]) or sample average approximation (see [83]). These two approaches are detailed in the coming two subsections. When, Y is a stochastic process, the parametric representation can be coupled with multilevel Monte Carlo as explained in Section 1.1.3.

1.1.1 Stochastic approximation

The idea is to build the sequence

$$\theta_{n+1} = \theta_n - \gamma_{n+1} H(\theta_n, Y_{n+1}) + \text{"correction"} \tag{1.1}$$

where $(Y_i)_i$ is an i.id. sequence following the law of Y and $(\gamma_n)_n$ is a positive decreasing sequence. In [L-15, L-7], we studied the asymptotic properties of the sequence $(\theta_n)_n$ (convergence and central limit theorem). A general framework for the study of the adaptive estimator of $\mathbb{E}[f(X)]$, defined by $\frac{1}{n}\sum_{i=1}^n h(\theta_{i-1}, Y_i)$, was developed in [L-11], which was a joint work with Bernard Lapeyre.

Chapter 2 presents the results obtained on stochastic approximation.

1.1.2 The *sample average approximation* approach

Instead of using stochastic approximation, one can apply deterministic optimisation techniques on sample averages. The second moment v is approximated by $v_n(\theta) = \frac{1}{n}\sum_{i=1} h(\theta, Y_i)^2$ where the sequence $(Y_i)_i$ is i.i.d. according to the law of Y. Then, we compute $\theta_n = \arg\min_\theta v_n(\theta)$ by a deterministic optimisation method and we build $M_n = \frac{1}{n}\sum_{i=1} h(\theta_n, \bar{Y}_i)$ where the sequence $(\bar{Y}_i)_i$ is i.i.d. following the law of Y.

Two cases have been studied.

- In [L-12], Benjamin Jourdain and I considered the case where $\bar{Y}_i = Y_i$ for all i and Y is a Gaussian vector.
- In [L-5], we considered the case where Y has a Gaussian part and a Poisson part. In this work, we assumed that the sequences $(Y_i)_i$ and $(\bar{Y}_i)_i$ were independent.

The algorithms developed in [L-12, L-5] and summarized in Chapter 3 are currently used by Mentor Graphics and Natixis. While writing this chapter, I managed to improve some of the results originally published in [L-5] and I chose to present them in the light of new strong laws of large numbers for doubly indexed sequences (see Section 7.1), which were not available to us at the time we wrote the original article.

1.1.3 Coupling Multilevel Monte Carlo with importance sampling

In [L-1], we studied with Ahmed Kebaier how to couple multilevel Monte Carlo with importance sampling. This problem can fit in the framework of the above paragraph by considering the case in which the Gaussian vector Y consists of the Brownian increments used in the time discretization of an underlying process X. In this situation, a discretization error adds to the usual Monte Carlo error. The multilevel Monte Carlo method enables us to better balance these two errors. We have studied how to further improve the variance of the multilevel estimator by cleverly applying the technique developed in [L-12, L-5] and summarized in Chapter 3.

Our contribution to importance sampling for multilevel Monte Carlo is detailed in Chapter 4.

1.2 A stochastic optimization point of view to American options

Consider an optimal stopping problem with finite time horizon T, defined on a filtered probability space with Brownian filtration $\mathbb{F} = (\mathcal{F}_t)_{0\leq t\leq T}$,

$$U_0 = \sup_{\tau\in\mathcal{T}} \mathbb{E}[\psi(X_\tau)]$$

where $\mathcal{T}$ denotes the set of all $\mathbb{F}$ stopping times and X is a Markov process. This optimal stopping problem admits a dual formulation [48, 80]

$$U_0 = \inf_{Y\in L^2(\mathcal{F}_T)} \mathbb{E}\left[\sup_{0\leq t\leq T} (\psi(X_t) - \mathbb{E}[Y|\mathcal{F}_t])\right]. \tag{1.2}$$

The quality of the approximation depends on the parametrisation of the space of squared integrable random variables, which has to allow the computation of the conditional expectations using closed formulae. In [L-21], we proposed to use the vector space of truncated Wiener chaos to approximate $L^2(\mathcal{F}_T)$. Then, (1.2) can be approximated by a finite dimensional optimisation problem, whose solution converges to U_0 when both the degree and the order of the truncation tend to infinity. We have studied the impact of the different parameters on the convergence rate of the approximation. Then, we have proposed a parallel algorithm along with a C++ implementation with impressive scaling properties on high performance computers.

This paper is summarized in Chapter 5.

1.3 Stochastic modelling for ferro–magnets

The behaviour of a magnetic moment μ submitted to an external field b is modeled by the Landau–Lifshitz equation

$$\frac{d\mu}{dt} = -\mu \wedge b - \alpha\mu \wedge (\mu \wedge b),$$

where $\alpha > 0$ and $\mu_0 \in S(\mathbb{R}^3)$. With S. Labbé, we have worked on the introduction of stochastic terms in this equation to take into account thermal effects.

- In [L-6], we proposed stochastic model preserving the norm of μ, which is a fundamental property of the physical system.

$$dY_t = -\mu_t \wedge (b\ dt + \varepsilon\ dW_t) - \alpha\mu_t \wedge (\mu_t \wedge (b\ dt + \varepsilon\ dW_t)) \text{ with } \mu_t = \frac{Y_t}{|Y_t|} \tag{1.3}$$

  where W is a Brownian motion in $\mathbb{R}^3$. In this model, μ converges to the unique stable equilibrium of the deterministic system and we determined the convergence rate in L^2.

- In [L-22], we studied other stochastic models and finally focused on

$$\partial\mu_t = -\mu_t \wedge (b\ dt + \varepsilon_t\ \partial W_t) - \alpha\mu_t \wedge (\mu_t \wedge (b\ dt + \varepsilon_t\ \partial W_t)) \tag{1.4}$$

  where the operator ∂ denotes the Stratonovich differential operator and $(\varepsilon_t)_t$ is a deterministic, positive and decreasing function. We showed that this stochastic system converges to $b/\left|b\right|$ and obtained an L^p convergence rate. Now, we work on the link between the function $(\varepsilon_t)_t$ and the decrease of the temperature.

Equations (1.3) and (1.4) can be seen as continuous time stochastic approximations. Therefore, it is not surprising that the techniques developed in these works remind us of the methods used in the beginning of Chapter 2.

These two papers are presented in Chapter 6.

1.4 Some key technical tools

In Chapter 7, we present several tools used at different places in the manuscript and which proved to be particularly efficient. I believe that they deserve some special emphasis as they may well be applied to other problems.

- Section 7.1 is dedicated to some strong laws of large numbers for doubly indexed sequences, which were proved in [L-1]. Based on these new laws, I managed to present some improved results in Section 3.3.

- My proof of the a.s. convergence of randomly truncated stochastic approximation (Chapter 2) relies on splitting the core of the martingale part from its remainder and then on carrying a pathwise deterministic analysis related to the ODE method (see [65]). The key ideas of this method further used in Chapter 6 are detailed in Section 7.2.

- My numerical experiments are all based on an open source scientific library PNL [L-19], which I have been developing for 10 years. I present in Section 7.3 my motivation for developing such a library and the key features I implemented in connection to my research activities.

1.5 High Performance Computing

Even though, I have finally decided not to write a dedicated chapter on High Performance Computing (HPC) in this document, I would like to emphasize this aspect in my research works as it has been playing a major role. Whenever I propose a new numerical method for high dimensional problems, I feel concerned with providing a scalable algorithm usable in real life problems. For instance, the algorithms developed in [L-8, L-1, L-21] were all successfully tested on clusters with several hundreds of processors. As a good parallel algorithm can hardly be obtained by paralleling a sequential algorithm, my will to propose HPC algorithms has often motivated my choices for one mathematical methodology rather than an other. Here, I describe two such situations

- In Chapter 5, the stochastic optimization problem coming from the dual formulation of the American option price is solved using *sample average approximation* rather than stochastic approximation. Stochastic approximation is inherently sequential, whereas sample average approximation enables us to use a parallel evaluation of the cost function, which basically writes as a Monte Carlo.

- In Chapter 4, we allowed for one importance sampling parameter per level instead of using a single importance sampling parameter for all the levels, which would have made all them dependent on one another as in [12]. In our approach, all the levels remain independent as if there were no importance sampling and can be solved in parallel, which is a key feature of multilevel Monte Carlo and makes it so popular.

All these HPC algorithms have been implemented in C++ with the help of PNL [L-19].

Chapter 2

Stochastic approximation

This chapter presents my theoretical results on the convergence of stochastic algorithms [L-15], [L-7] and an application to a general framework for adaptive Monte Carlo methods [L-11].

2.1 Introduction to stochastic approximation

The use of stochastic algorithms is widespread for solving stochastic optimization problems. These algorithms are extremely valuable for a practical use and particularly well suited to localize the zero of a function u. Such algorithms go back to the pioneering work of Robbins and Monro [79], who considered the sequence

$$X_{n+1} = X_n - \gamma_{n+1} u(X_n) - \gamma_{n+1} \delta M_{n+1} \tag{2.1}$$

to estimate the zero of the function u. The gain sequence $(\gamma_n)_n$ is a decreasing sequence of positive real numbers and $(\delta M_n)_n$ represents a random measurement error. Since their work, much attention has been drawn to the study of such recursive approximations. The first works were dealing with independent measurement error on the observations. A great effort has been made in this direction to weaken the conditions imposed on both the function u and the noise term δM_n. Using the ordinary differential equation technique, Kushner and Clark [65] proved a convergence result for a wider range of measurement noises and in particular for martingale increments.

Nevertheless, the assumptions required to ensure the convergence — basically, a sub-linear growth of u on average — are barely satisfied in practice, which dramatically reduces the range of applications. Chen and Zhu [29] proposed a modified algorithm to deal with fast growing functions. Their new algorithm can be written as

$$X_{n+1} = X_n - \gamma_{n+1} u(X_n) - \gamma_{n+1} \delta M_{n+1} + \gamma_{n+1} p_{n+1} \tag{2.2}$$

where $(p_n)_n$ is a truncation term ensuring that the sequence $(X_n)_n$ cannot jump too far ahead in a single move.

I studied both the almost sure convergence and the convergence rate of such algorithms. Several results already existed but the hypotheses considered differed quite significantly. The first result concerning the almost sure convergence can be found in [29]. The convergence was further studied in [30, 34]. All these studies were carried out using global assumptions on the noise, they basically required that the series $\sum_n \gamma_{n+1} \delta M_{n+1}$ converges a.s. Note that these properties are intimately linked to the paths of the algorithm itself, which makes them even harder to be checked in practice. Although, some results [28, 33] relying on local assumptions were already available, I aimed at giving a unified framework with no martingale structure for the noise terms δM_n.

2.2 Stochastic approximation with expanding truncations

Consider the problem of finding the root of a continuous function $u\colon x \in \mathbb{R}^d \longmapsto u(x) \in \mathbb{R}^d$, which is not easily tractable. We assume that we can only access u up to a measurement error embodied in the following by the sequence $(\delta M_n)_n$ and that the norm $|u(x)|^2$ grows faster than $|x|^2$ such that the standard Robbins-Monro algorithm (see (2.1)) fails. Instead, we consider the alternative procedure introduced by [29]. This technique consists in forcing the algorithm to remain in an increasing sequence of compact sets $(K_j)_j$ such that

$$\bigcup_{j=0}^{\infty} K_j \;=\; \mathbb{R}^d \quad \text{and} \quad \forall j,\; K_j \subsetneq \operatorname{int}(K_{j+1})\,. \tag{2.3}$$

It prevents the algorithm from blowing up during the first iterates. The general idea is to monitor the moves of the dynamical system (2.1) and to pull the dynamics back to a fixed compact set when the algorithm makes too big steps. This mechanism ensures the stability of the algorithm — the existence of a recurrent set — and is definitely essential to prove the convergence.

We introduce a bounded sequence of random vectors $(Y_n)_n$ with values in $\mathbb{R}^d$, which represents the reinitialization values. Then, it is natural to impose that for all n, $Y_n \in K_n$; actually the boundedness of Y_n ensures the existence of a deterministic integer N such that for all $n \geq N$, $Y_n \in K_N$. Let $(\gamma_n)_n$ be a decreasing sequence of positive real numbers satisfying $\sum_n \gamma_n = \infty$ and $\sum_n \gamma_n^2 < \infty$. For $X_0 \in \mathbb{R}^d$ and $\sigma_0 = 0$, we define the sequences of random variables $(X_n)_n$ and $(\sigma_n)_n$ by

$$\begin{cases} X_{n+\frac{1}{2}} = X_n - \gamma_{n+1} u(X_n) - \gamma_{n+1}\delta M_{n+1}, & \\ X_{n+1} = X_{n+\frac{1}{2}} \qquad \text{and} \quad \sigma_{n+1} = \sigma_n, & \text{if } X_{n+\frac{1}{2}} \in K_{\sigma_n} \\ X_{n+1} = Y_{\sigma_n} \qquad \text{and} \quad \sigma_{n+1} = \sigma_n + 1, & \text{if } X_{n+\frac{1}{2}} \notin K_{\sigma_n}. \end{cases} \tag{2.4}$$

Let $\mathbb{F} = (\mathcal{F}_n)_n$ be the $\sigma-$algebra generated by the noise terms, for all n, $\mathcal{F}_n = \sigma(\delta M_k, k \leq n)$. To ensure that the sequence $(X_n)_n$ is $\mathbb{F}-$adapted, we impose that the sequence $(Y_n)_n$ is also $\mathbb{F}-$adapted. Note that Y_n can in particular be any bounded measurable function of $(X_0, X_1, \cdots, X_n)$.

It is more convenient to rewrite Equation (2.4) as follows

$$X_{n+1} = X_n - \gamma_{n+1} u(X_n) - \gamma_{n+1}\delta M_{n+1} + \gamma_{n+1} p_{n+1} \tag{2.5}$$

where

$$p_{n+1} = \left(u(X_n) + \delta M_{n+1} + \frac{1}{\gamma_{n+1}}(Y_{\sigma_n} - X_n) \right) \mathbf{1}_{X_{n+\frac{1}{2}} \notin K_{\sigma_n}}.$$

2.2.1 Hypotheses

To study the a.s. convergence and the convergence rate of $(X_n)_n$ defined by (2.5), we introduce the following assumptions

($\mathcal{H}$2.1)

i. The function u is continuous.
$\exists x_\star \in \mathbb{R}^d$ s.t. $u(x_\star) = 0$ and $\forall x \in \mathbb{R}^d,\; x \neq x_\star,\; (x - x_\star)\cdot u(x) > 0$.

ii. There exist a function $y : \mathbb{R}^d \to \mathcal{M}_{d\times d}$ satisfying $\lim_{|x|\to 0} |y(x)| = 0$ and a repulsive[1] matrix $A \in \mathcal{M}_{d\times d}$ such that

$$u(x) = A(x - x_\star) + y(x - x_\star)(x - x_\star).$$

[1] A matrix is said to be repulsive if all its eigenvalues have positive real parts.

($\mathcal{H}$2.2) For any $q > 0$, the series $\sum_n \gamma_{n+1}\delta M_{n+1}\mathbf{1}_{|X_n - x^\star|\leq q}$ converges almost surely.

($\mathcal{H}$2.3) The sequence $(\delta M_n)_n$ is a sequence of martingale increments, ie. $\mathbb{E}[\delta M_{n+1}|\mathcal{F}_n] = 0$ and moreover

i. there exist two real numbers $\rho > 0$ and $\eta > 0$ such that

$$\kappa = \sup_n \mathbb{E}\left(|\delta M_n|^{2+\rho}\,\mathbf{1}_{|X_{n-1}-x_\star|\leq\eta}\right) < \infty;$$

ii. there exists a symmetric positive definite matrix $\Sigma \in \mathcal{M}_{d\times d}$ such that

$$\mathbb{E}\left(\delta M_n \delta M_n^T \big| \mathcal{F}_{n-1}\right)\mathbf{1}_{|X_{n-1}-x_\star|\leq\eta} \xrightarrow[n\to\infty]{\mathbb{P}} \Sigma.$$

($\mathcal{H}$2.4) There exists $\mu > 0$ such that $\forall n \geq 0,\ d(x_\star, \partial K_n) \geq \mu$.

Before investigating the long time behaviour of $(X_n)_n$, it is worth having a closer look at these assumptions to better understand the range of applicability.

- Hypothesis ($\mathcal{H}$2.1-*ii)* is equivalent to saying that u is differentiable at $x_\star$. The Hypothesis ($\mathcal{H}$2.1-*i)* is satisfied as soon as u can be interpreted as the gradient of a strictly convex function and in this case the matrix A is the Hessian matrix at $x_\star$.

- In [28], ($\mathcal{H}$2.2) was replaced by a condition on the convergent subsequences of the algorithm. We believe our assumption is easier to check in practice as it basically boils down to proving the local a.s. convergence of a martingale. In Section 2.2.3, we provide a sufficient and simple condition to ensure ($\mathcal{H}$2.2) when the function u is given as an expectation.

- Hypothesis ($\mathcal{H}$2.3-*i)* corresponds to some local uniform integrability condition and reminds us of Lindeberg's condition. ($\mathcal{H}$2.3-*ii)* guaranties the convergence of the angle bracket of the martingale and is key to obtain a central limit theorem. Assumption ($\mathcal{H}$2.3) is only required to study the convergence rate of $(X_n)_n$.

- Hypothesis ($\mathcal{H}$2.4) is only required for technical reasons but one does not need to be concerned with it in practice. It reminds us of the case of constrained stochastic algorithms for which the central limit theorem can only be proved for non saturated constraints.

2.2.2 Main results

Almost sure convergence For the a.s. convergence, no particular structure of the noise term is required, in particular we do not need any martingale assumption. Most of the time, the sequence $(\gamma_n)_n$ is assumed to be deterministic or at least predictable whereas our proof can cope with anticipative random sequences $(\gamma_n)_n$. Obviously, it does not make sense to consider an anticipative gain sequence but we can handle the case where (γ_n) is $\mathbb{F}$-adapted but not $\mathbb{F}-$predictable. Note that when $(\gamma_n)_n$ is not predictable, assuming that $(\delta M_n)_n$ are martingale increments does not ensure anymore that $(\sum_{k=1}^n \gamma_k \delta M_k)_n$ is still a martingale.

Theorem 2.2.1 *Assume $\gamma_n \downarrow 0$, $\sum_n \gamma_n = \infty$. Under Hypothesis ($\mathcal{H}$2.1-i) and ($\mathcal{H}$2.2), the sequence $(X_n)_n$ defined by* (2.5) *converges a.s. to $x_\star$ and moreover the sequence $(\sigma_n)_n$ is a.s. finite.*

Here, we only present the main ideas sustaining the proof of Theorem 2.2.1. They key tool we developed to study a dynamical system as (2.5) is to split the core of the martingale from its remainder in order to introduce an auxiliary system with no extra *random* term. See Section 7.2 for a detailed presentation of the method.

Sketch of the proof (Theorem 2.2.1). Assume we have already proved that $\sup_n \sigma_n < \infty$ a.s. and there a.s. exists $q > 0$ s.t. for large enough $n \geq 0$, $|X_n - x_\star|^2 \leq q$. Therefore, we deduce from ($\mathcal{H}$2.2) that $\sum_n \gamma_{n+1}\delta M_{n+1}$ converges a.s. Consider the auxiliary sequence $(X'_n)_n$ defined by

$$X'_n = X_n - \sum_{i=n+1}^{\infty} \gamma_i \delta M_i.$$

Let $\varepsilon > 0$. There exists N s.t. for all $n \geq N$, $\left|\sum_{i=n+1}^{\infty} \gamma_i \delta M_i\right| \leq \varepsilon$ and $(\sigma_n)_{n\geq N}$ is constant. For $n \geq N$, $X'_{n+1} = X'_n - \gamma_{n+1} u(X_n)$. Choosing $\varepsilon > 1$ guaranties that for all $n \geq N$, $|X'_n - x_\star|^2 \leq q + 1$. Let $\bar{u} \stackrel{\Delta}{=} \sup_{|x-x_\star|^2 \leq q+1} |u(x)|$.

$$\begin{aligned}
\left|X'_{n+1} - x_\star\right|^2 &\leq \left|X'_n - x_\star\right|^2 - 2\gamma_{n+1}(X'_n - x_\star)\cdot u(X_n) + \gamma_{n+1}^2 \left|u(X_n)\right|^2, \\
&\leq \left|X'_n - x_\star\right|^2 - 2\gamma_{n+1}(X_n - x_\star)\cdot u(X_n) + \gamma_{n+1}^2 \bar{u}^2 + 2\gamma_{n+1}\bar{u}\varepsilon.
\end{aligned}$$

Let $\delta > 0, \delta > \varepsilon$. If $|X'_n - x_\star| > 2\delta$, then $|X_n - x_\star| > 2\delta - \varepsilon > \delta$. We know from Hypothesis ($\mathcal{H}$2.1) that $\inf_{|x-x_\star|>\delta}(x - x_\star \mid u(x)) \geq c > 0$. Hence,

$$\begin{aligned}
\left|X'_{n+1} - x_\star\right|^2 &\leq \left|X'_n - x_\star\right|^2 - 2\gamma_{n+1}(X'_n - x_\star)\cdot u(X_n) + \gamma_{n+1}^2 \left|u(X_n)\right|^2, \\
&\leq \left|X'_n - x_\star\right|^2 - 2\gamma_{n+1}(c\mathbf{1}_{|X'_n - x_\star|>2\delta} - \bar{u}\varepsilon) + \gamma_{n+1}^2 \bar{u}^2.
\end{aligned}$$

The integer N can be chosen large enough s.t. $2(c - \varepsilon\bar{u}) - \gamma_{n+1}\bar{u}^2 > 0$. Since $\sum_n \gamma_n = \infty$, each time $|X'_n - x_\star| > 2\delta$, the sequence X'_n is driven back into the ball $\bar{B}(x_\star, \delta)$ in a finite number of steps. Moreover, $\lim_{n\to\infty} 2\gamma_{n+1}\bar{u}\varepsilon + \gamma_{n+1}^2\bar{u}^2 = 0$. Hence, $\limsup_n |X'_n - x_\star|^2 \leq 4\delta^2$ for all $\delta > 0$. This proves that $X'_n \to x_\star$, which in turn implies that $X_n \to x_\star$. ■

Central limit theorem A central limit theorem can be obtained by centering the iterates around their limit and applying the rescaling factor $(\gamma_n)^{-1/2}$. An earlier version of the central limit theorem could be found in [28] but some arguments of his proof deserved to be developed to make it crystal clear and Chen could not afford the term $\mathbf{1}_{|X_{n-1}-x_\star|\leq\eta}$, which is essential for Section 2.2.3.

Theorem 2.2.2 *Consider sequences $(\gamma_n)_n$ of the form $\gamma_n = \frac{\gamma}{(n+1)^\alpha}$, with $1/2 < \alpha \leq 1$. Assume Hypotheses ($\mathcal{H}$2.1) to ($\mathcal{H}$2.4) and one the following conditions*

- *If $\alpha = 1$ and the matrix $\frac{1}{2\gamma}I - A$ is stable, set $Q = A - I/(2\gamma)$;*
- *If $1/2 < \alpha < 1$, set $Q = A$.*

Then, the sequence $\left(n^{\alpha/2}(X_n - x_\star)\right)_n$ converges in distribution to a normal random variable with mean 0 and covariance matrix

$$V = \gamma \int_0^\infty \mathrm{e}^{-Qt}\, \Sigma\, \mathrm{e}^{-Q^T t}\, dt,$$

which solves $QV + VQ^T - \Sigma = 0$.

The proof of this result being quite technical, we refer the reader to [L-7]. The key tool we developed was the introduction of the collection of sets

$$A_n = \left\{ \sup_{n \geq m \geq N_0} |X_m - x_\star| \leq \eta \right\}$$

satisfying that for every $1 > \varepsilon > 0$ and $\eta > 0$, there exists N_0, s.t. $\mathbb{P}(A_n) \geq 1 - \varepsilon$ for all $n > N_0$. Then, we only needed to prove a "local" central limit theorem for the sequence $(X_n - x_\star)\mathbf{1}_{A_n}$. The idea of introducing subsets of Ω with probabilities increasing to 1 to obtain central limit theorems is used again in Chapters 3 and 4.

In the case $\alpha = 1$, the assumption on the repulsive behaviour of $\gamma A - \frac{1}{2} I$ imposes to choose γ sufficiently large in order to obtain a central limit theorem. Actually, we can afford more general step sequences and consider $\gamma_n = \frac{\Gamma}{n+1}$ where Γ is a $d \times d$ invertible matrix. The matrix Γ can be absorbed into the function u and the sequence $(\delta M_n)_n$ such that A becomes ΓA and Σ becomes $\Gamma \Sigma \Gamma^T$. As for now, we assume that the assumptions of Theorem 2.2.2 are still valid after absorbing Γ in such a way. Hence, the central limit theorem writes

$$\sqrt{n}(X_n - x_\star) \xrightarrow[n \to \infty]{\mathcal{D}} \mathcal{N}(0, V_\Gamma)$$

where

$$V_\Gamma = \int_0^\infty \mathrm{e}^{(I/2 - \Gamma A)t} \, \Gamma \Sigma \Gamma^T \, \mathrm{e}^{\left(I/2 - A^T \Gamma^T\right)t} \, dt.$$

Clearly, we would like to choose Γ *minimizing* the asymptotic variance, a possible criterion could be to minimize $\mathrm{tr}(V_\Gamma)$, which corresponds to the sum of the marginal variances.

$$\begin{aligned}
\mathrm{tr}(V_\Gamma) &= \mathrm{tr}\left(\int_0^\infty \mathrm{e}^{(I/2 - \Gamma A)t} \, \mathrm{e}^{\left(I/2 - A^T \Gamma^T\right)t} \, \Gamma \Sigma \Gamma^T dt \right) \\
&= \mathrm{tr}\left(\int_0^\infty \mathrm{e}^{\left(I - \Gamma A - A^T \Gamma^T\right)t} \, dt \; \Gamma \Sigma \Gamma^T \right) \\
&= \mathrm{tr}\left((I - \Gamma A - A^T \Gamma^T)^{-1} \Gamma \Sigma \Gamma^T \right).
\end{aligned}$$

If we compute the differential form of the application $\Gamma \mapsto \mathrm{tr}(V_\Gamma)$ applied to the matrix H we find after grouping terms

$$2 \, \mathrm{tr}\left(-H^T A (I - \Gamma A - A^T \Gamma^T)^{-1} \Gamma \Sigma \Gamma^T (I - \Gamma A - A^T \Gamma^T)^{-1} + H \Sigma \Gamma^T (I - \Gamma A - A^T \Gamma^T)^{-1} \right).$$

This quantity is zero for all H if and only if

$$A(I - \Gamma A - A^T \Gamma^T)^{-1} \Gamma + I = 0.$$

Then, we deduce that $\mathrm{tr}(V_\Gamma)$ is minimum for $\Gamma = A^{-1}$ and the asymptotically optimal covariance is

$$V_{A^{-1}} = A^{-1} \Sigma (A^{-1})^T. \tag{2.6}$$

2.2.3 A standard noise structure

In practical applications, the function u is very often defined as an expectation $u(x) = \mathbb{E}[U(x, Z)]$ where Z is a random variable with values in $\mathbb{R}^m$ and one only has access to samples from $U(x, Z)$. In such a framework, the stochastic approximation writes

$$X_{n+1} = X_n - \gamma_{n+1} U(X_n, Z_{n+1})$$

where the Z_n's are i.i.d according to the distribution of Z. This equation naturally fits in the framework defined (2.4) when choosing $\delta M_{n+1} = U(X_n, Z_{n+1}) - u(X_n)$. Let $\mathbb{F}$ be the filtration generated by the Z_n's, then $\mathbb{E}[\delta M_{n+1}|\mathcal{F}_n] = 0$. In this section, we assume that the sequence $(\gamma_n)_n$ is predictable.

Let $q > 0$, we define the sequence $(M_n^q)_n$ by $M_n^q = \sum_{i=1}^n \gamma_i \delta M_i \mathbf{1}_{|X_{i-1}-x_\star|\leq q}$, which is clearly a $\mathbb{F}$-martingale with angle bracket

$$\begin{aligned}\langle M^q\rangle_n &= \sum_{i=1}^n \gamma_i^2 \mathbb{E}[\delta M_i \delta M_i^T|\mathcal{F}_{i-1}]\mathbf{1}_{|X_{i-1}-x_\star|\leq q},\\ &= \sum_{i=0}^{n-1} \gamma_i^2 (\mathbb{E}[U(X_i, Z_{i+1})U(X_i, Z_{i+1})^T|\,\mathcal{F}_i] - u(X_i)u(X_i)^T)\mathbf{1}_{|X_{i-1}-x_\star|\leq q}.\end{aligned}$$

If $\sum_n \gamma_n^2 < \infty$ and the function $x \longmapsto \mathbb{E}[|U(x,Z)|^2]$ is bounded on any compact set, we immediately deduce that $\sup_n \langle M^q\rangle_n < \infty$ a.s. Then, the strong law of large numbers for square integrable martingales (see for instance [46, 71]) yields the a.s. convergence of M_n^q when $n \to \infty$ and therefore ($\mathcal{H}2.2$) is satisfied. Consequently, Theorem 2.2.1 takes a much simpler form

Corollary 2.2.3 *Assume $\sum_n \gamma_n = \infty$ and $\sum_n \gamma_n^2 < \infty$ a.s. Under Hypothesis ($\mathcal{H}2.1$-i) and if the function $x \longmapsto \mathbb{E}[|U(x,Z)|^2]$ is locally bounded, then the sequence $(X_n)_n$ defined by* (2.5) *converges a.s. to $x_\star$ and moreover the sequence $(\sigma_n)_n$ is a.s. finite.*

The assumption ($\mathcal{H}2.3$) required to obtain a central limit theorem can also be reformulated in a much simpler way. Condition ($\mathcal{H}2.3$-*i)* is satisfied as soon as the function $x \longmapsto \mathbb{E}[|U(x,Z)|^{2+\rho}]$ is locally bounded and u is continuous. Moreover, if the function $x \longmapsto \mathbb{E}[U(x,Z)U(x,Z)^T]$ is continuous at $x_\star$, Assumption ($\mathcal{H}2.3$-*ii)* holds with $\Sigma = \mathrm{Cov}(U(x_\star, Z))$.

Corollary 2.2.4 *Consider sequences $(\gamma_n)_n$ of the form $\gamma_n = \frac{\gamma}{(n+1)^\alpha}$, with $1/2 < \alpha \leq 1$. Assume Hypotheses ($\mathcal{H}2.1$-i) and ($\mathcal{H}2.4$). Moreover, assume u is continuous, the function $x \longmapsto \mathbb{E}[|U(x,Z)|^{2+\rho}]$ is locally bounded for some $\rho > 0$ and the function $x \longmapsto \mathbb{E}[U(x,Z)U(x,Z)^T]$ is continuous at $x_\star$. Then, the conclusion of Theorem 2.2.2 holds.*

All the delicate assumptions involving $(\delta M_n)_n$ take a much simpler form in this context as we managed to narrow the required conditions to compact sets. These new assumptions are far easier to check in practical applications and make the extra difficulty of the proofs worth it.

2.2.4 Averaging the iterates

When it comes to practically using stochastic approximation, people often implement an averaging principle on top of it to smooth the convergence. It is based on the idea that if a sequence converges, its Césaro mean converges more smoothly. Typically, we propose the following companion algorithm to (2.5)

$$\bar{X}_n = \frac{1}{n-m+1}\sum_{i=m}^n X_i$$

where m is the starting index of the averaging process. Clearly, when $(X_n)_n$ converges so does $(\bar{X}_n)_n$. In practice, we would like $(\bar{X}_n)_n$ to converge faster and more smoothly, which makes the choice of m not so easy. If m is too large, there is no smoothing process and if m too small, it takes ages to forget the initial condition and therefore makes the convergence slower. At this point, it becomes clear that

m should depend on n. Consider sequences $(\gamma_n)_n$ of the form $\gamma_n = \frac{\gamma}{(n+1)^\alpha}$, with $1/2 < \alpha < 1$ and a moving window average

$$\tilde{X}_n(\tau) = \frac{1}{n - m_n + 1} \sum_{i=m_n}^{n} X_i \tag{2.7}$$

with $m_n = \sup\{k \geq 1 \ : \ k + \tau k^\alpha \leq n\} \wedge n$ and $\tau > 0$ being a real parameter monitoring the length of the averaging window. For any $\tau > 0$, the convergence of $(\tilde{X}_n(\tau))_n$ easily ensues from Theorem 2.2.1 or Corollary 2.2.3. The asymptotic normality has been studied in [L-20].

Theorem 2.2.5 . *Assume Hypotheses ($\mathcal{H}$2.1) to ($\mathcal{H}$2.4). Then, for any $\tau > 0$, the sequence $\left(n^{\alpha/2}(\tilde{X}_n(\tau) - x_\star)\right)_n$ converges in distribution to a normal random vector with mean 0 and covariance matrix*

$$\tilde{V}(\tau) = \frac{A^{-1}\Sigma A}{\tau} + O(\tau^{-2}).$$

Note that the zero order approximation of $\tau\tilde{V}(\tau)$ corresponds to the asymptotic variance obtained with the best possible matrix valued gain sequence (see the discussion after Theorem 2.2.2 and Equation (2.6)). In practice, it is hardly feasible to use the best gain sequence as it requires to know the minimizer $x_\star$. The averaging process presented in this paragraph enables us to almost reach the asymptotically optimal variance and makes the numerical convergence of the algorithm much smoother and more robust to the choice of γ.

2.3 Application to adaptive Monte Carlo methods

2.3.1 A general framework

Consider the problem of computing the expectation $\mathbb{E}[Y]$ of a real valued random variable Y and assume Y admits a parametric representation such that

$$\mathbb{E}[Y] = \mathbb{E}[H(x, Z)] \quad \text{for all } x \in \mathbb{R}^d, \tag{2.8}$$

where Z is a random vector with values in $\mathbb{R}^m$ and $H : \mathbb{R}^d \times \mathbb{R}^m \longmapsto \mathbb{R}$ is a measurable function satisfying $\mathbb{E}\,|H(x, Z)| < \infty$ for all $x \in \mathbb{R}^d$. We also impose that

$$x \longmapsto v(x) = \mathrm{Var}(H(x, Z)) \text{ is finite for all } x \in \mathbb{R}^d, \tag{2.9}$$

We want to make the most of this free parameter x to settle an automatic variance reduction method, see [55] for a recent survey on adaptive variance reduction. It consists in first finding a minimiser $x_\star$ of the variance v and then plugging it into a Monte Carlo method with a narrower confidence interval. This technique heavily relies on the ability to find a parametric representation and to effectively minimize the function v. Many papers have been written on how to construct parametric representations $H(x, Z)$ for several kinds of random variables Z. We mainly have in mind examples based on control variates (see [44, 61, 62]) or importance sampling (see [3, 4, 86]).

Assume we have a parametric representation of the form $H(x, Z)$ satisfying Equations (2.8) and (2.9). Let $(Z_n)_n$ be an i.i.d. sequence of random vectors with the distribution of Z. Assume we know how to use the sequence $(Z_n)_n$ to build an estimator X_n of $x_\star$ adapted to the filtration $\mathbb{F} = (\mathcal{F}_n = \sigma(Z_1, \ldots, Z_n))_n$. Once such an approximation is available, there are at least two ways of using it to devise a variance reduction method.

The non-adaptive algorithm

Algorithm 2.3.1 (Non adaptive importance sampling (NADIS)) *Let n be the number of samples used for the Monte Carlo computation. Draw a second set of n samples $(Z'_1, \dots, Z'_n)$ independent of $(Z_1, \dots, Z_n)$ and compute*

$$\bar{\xi}_n = \frac{1}{n} \sum_{i=1}^{n} H(X_n, Z'_i).$$

This algorithm has been studied in [4, 86] and required $2n$ samples. It may use less than $2n$ samples if the estimation of $x_\star$ is performed on a smaller number of samples but then it raises the question of how many samples to use.

The adaptive algorithm The adaptive approach is to use the *same samples* $(X_1, \dots, X_n)$ to compute X_n and the Monte Carlo estimator. Compared to the sequential algorithm, the adaptive one uses half of the samples.

Algorithm 2.3.2 (Adaptive Importance Sampling (ADIS)) *Let n be the number of samples used for the Monte Carlo computation.*
For X_0 fixed in $\mathbb{R}^d$, compute

$$\xi_n = \frac{1}{n} \sum_{i=1}^{n} H(X_{i-1}, Z_i). \tag{2.10}$$

The sequence $(\xi_i)_i$ can be written in a recursive manner so that it can be updated online each time a new iterate X_i is drawn

$$\xi_{i+1} = \frac{i}{i+1} \xi_i + \frac{1}{i+1} H(X_i, Z_{i+1}), \quad \text{with } \xi_0 = 0.$$

The online update of the sequence $(\xi_i)_i$ enables us to avoid storing the whole sequence $(Z_1, \dots, Z_n)$ for computing ξ_n. This adaptive algorithm was first studied in [4] under assumptions to be verified along the path of $(X_n)_n$, which makes them hard to check in practise. In [L-11], we have proved a new convergence result under local integrability conditions on the function H, see Theorem 2.3.3.

2.3.1.1 Computing the optimal parameter

In this section, we are interested in effective ways to compute or at least approximate the optimal parameter $x_\star$. To do so, we further assume that the function v is strictly convex, goes to infinity at infinity and is continuously differentiable. Moreover, we suppose that ∇v admits a representation as an expectation

$$\nabla v(x) = \mathbb{E}[U(x, Z)].$$

When the function H is sufficiently smooth, $U(x, Z) = 2H(x, Z)\nabla_x H(x, Z)$ a.s. for all $x \in \mathbb{R}^d$. Then, the optimal parameter $x_\star$ is uniquely determined by the equation $\mathbb{E}[U(x_\star, Z)] = 0$, which fits exactly in the framework of Section 2.2.3. Define the sequence $(X_n)_n$ as in (2.4)

$$X_{n+1} = \mathcal{T}_{K_{\sigma_n}} \left(X_n - \gamma_{n+1} U(X_n, Z_{n+1}) \right)$$

where $\mathcal{T}_{K_{\sigma_n}}$ denotes the truncation on the compact set K_{σ_n}. Corollary 2.2.3 yields the convergence of the sequence $(X_n)_n$ to $x_\star$.

2.3.2 Convergence of the adaptive Monte Carlo method

An adaptive strong law of large numbers

Theorem 2.3.3 (Adaptive strong law of large numbers) *Assume Equation* (2.8) *and* (2.9) *hold. Let* $(X_n)_{n\geq 0}$ *be a* $(\mathcal{F}_n)-$*adapted sequence with values in* $\mathbb{R}^d$ *such that for all* $n \geq 0$, $X_n < \infty$ *a.s and for any compact subset* $K \subset \mathbb{R}^d$, $\sup_{x\in K} \mathbb{E}(|H(x,Z)|^2) < \infty$. *If*

$$\inf_{x\in\mathbb{R}^d} v(x) > 0 \qquad \textit{and} \qquad \frac{1}{n}\sum_{k=0}^{n} v(X_k) < \infty \quad \textit{a.s.}, \tag{2.11}$$

then ξ_n *converges a.s. to* $\mathbb{E}(Y)$.

Proof. For any $p \geq 0$, we define $\tau_p = \inf\{k \geq 0; |X_k| \geq p\}$. The sequence $(\tau_p)_p$ is an increasing sequence of $\mathbb{F}-$stopping times such that $\lim_{p\to\infty} \tau_p \uparrow \infty$ a.s.. Let $M_n = \sum_{i=0}^{n-1} H(X_i, Z_{i+1}) - \mathbb{E}(Y)$. We introduce $M_n^{\tau_p} = M_{\tau_p \wedge n}$ defined by

$$M_n^{\tau_p} = \sum_{i=0}^{n-1\wedge\tau_p} H(X_i, Z_{i+1}) - \mathbb{E}(Y) = \sum_{i=0}^{n-1} (H(X_i, Z_{i+1}) - \mathbb{E}(Y))\mathbf{1}_{i\leq\tau_p}.$$

$\mathbb{E}(|H(X_i, Z_{i+1}) - \mathbb{E}(Y)|^2 \mathbf{1}_{i\leq\tau_p}) \leq \mathbb{E}(\mathbf{1}_{i\leq\tau_p}\mathbb{E}(|H(X,Z) - \mathbb{E}(Y)|^2)_{X=X_i})$. On the set $\{i \leq \tau_p\}$, the conditional expectation is bounded from above by $\sup_{|X|\leq p} v(X)$. Hence, the sequence $(M_n^{\tau_p})_n$ is square integrable and it is obvious that $(M_n^{\tau_p})_n$ is a martingale, which means that the sequence $(M_n)_n$ is a locally square integrable martingale (i.e. a local martingale which is locally square integrable).

$$\langle M\rangle_n = \sum_{i=0}^{n-1} \mathbb{E}((H(X_i, Z_{i+1}) - \mathbb{E}(Y))^2 | \mathcal{F}_i) = \sum_{i=0}^{n-1} v(X_i).$$

By Condition (2.11), we have a.s. $\limsup_n \frac{1}{n}\langle M\rangle_n < \infty$ and $\liminf_n \frac{1}{n}\langle M\rangle_n > 0$. The strong law of large numbers for locally $\mathbb{L}^2$ martingales (see [71]) yields the result. ■

The sequence $(X_n)_n$ can be any sequence adapted to $(Z_n)_{n\geq 1}$ convergent or not. For instance, $(X_n)_n$ can be an ergodic Markov chain distributed around the minimizer $x_\star$ such as Monte Carlo Markov Chain algorithms.

Remark 2.3.4 When the sequence $(X_n)_{n\geq 0}$ converges a.s. to a deterministic constant x_∞, it is sufficient to assume that v is continuous at x_∞ and $v(x_\infty) > 0$ to ensure that (2.11) is satisfied. There is no need to impose that $x_\infty = x_\star$ although it is undoubtedly wished in practice.

A Central limit theorem for the adaptive strong law of large numbers To derive a central limit theorem for the adaptive estimator ξ_n, we need a central limit theorem for locally square integrable martingales, whose convergence rate has been extensively studied. We refer to the works of Rebolledo [77], Jacod and Shiryaev [53], Hall and Heyde [46] and Whitt [90] to find different statements of central limit theorems for locally square integrable càdlàg martingales in continuous time, from which theorems can easily be deduced for discrete time locally square integrable martingales.

Theorem 2.3.5 *Assume Equation* (2.8) *and* (2.9) *hold. Let* $(X_n)_{n\geq 0}$ *be a* $\mathbb{F}-$*adapted sequence with values in* $\mathbb{R}^d$ *such that for all* $n \geq 0$, $X_n < \infty$ *a.s and converging to some deterministic value* x_∞. *Assume there exists* $\eta > 0$ *such that the function* $s_{2+\eta} : x \in \mathbb{R}^d \longmapsto \mathbb{E}\left(|H(x,Z)|^{2+\eta}\right)$ *is finite for all* $x \in \mathbb{R}^d$ *and continuous at* x_∞. *Moreover, if* v *is continuous at* x_∞ *and* $v(x_\infty) > 0$, *then,* $\sqrt{n}(\xi_n - \mathbb{E}(Y)) \xrightarrow{law} \mathcal{N}(0, v(x_\infty))$.

Proof. We know from the proof of Theorem 2.3.3 that $M_n = \sum_{i=0}^{n-1} H(X_i, Z_{i+1}) - \mathbb{E}(Y)$ is a locally square integrable martingale and that $\frac{1}{n}\langle M\rangle_n$ converges a.s. to $v(x_\infty)$.

$$\frac{1}{n}\sum_{i=0}^{n-1} \mathbb{E}(|H(X_i, Z_{i+1}) - \mathbb{E}(Y)|^{2+\eta}|\mathcal{F}_i) \leq c\left(\frac{1}{n}\sum_{i=0}^{n-1} s_{2+\eta}(X_i) + \mathbb{E}(Y)^{2+\eta}\right).$$

The term on the r.h.s is bounded thanks to the continuity of $s_{2+\eta}$ at x_∞. Hence, the local martingale $(M_n)_n$ satisfies Lindeberg's condition. The result ensues from the central limit theorem for locally $\mathbb{L}^2$ martingales. ■

Remark 2.3.6 None of the assumptions of Theorems 2.3.3 and 2.3.5 are to be checked along the path of $(X_n)_n$, which makes these results valuable in practice. This improvement was possible without adding integrability conditions on $\sup_{x\in K}|H(x, Z)|$ for any compact sets K by relying on the theory on locally square integrable martingales and the asymptotic normality of ξ_n is proved under with very light conditions, namely some integrability properties on $|H(x, Z)|$ for any *fixed* x.

From a practical point of view, ti is desirable to have a central limit theorem using an estimator of the limiting variance. In a crude Monte Carlo setting, it is sufficient to rescale the estimator by an estimator of its variance after centering it. In our setting, the estimator is again obtained by martingale arguments.

Corollary 2.3.7 *Assume Equation* (2.8) *and* (2.9) *hold. Let* $(X_n)_{n\geq 0}$ *be a* $\mathcal{F}_n-$*adapted sequence with values in* $\mathbb{R}^d$ *such that for all* $n \geq 0$, $X_n < \infty$ *a.s and converging to some deterministic value* x_∞. *Assume there exists* $\eta > 0$ *such that the function* $s_{4+\eta} : x \in \mathbb{R}^d \longmapsto \mathbb{E}\left(|H(x, Z)|^{4+\eta}\right)$ *is finite for all* $x \in \mathbb{R}^d$ *and continuous at* x_∞. *Then,* $v_n^2 = \frac{1}{n}\sum_{i=0}^{n-1} H(X_i, Z_{i+1})^2 - \xi_n^2 \xrightarrow{a.s.} v(x_\infty)$. *If moreover* $v(x_\infty) > 0$, *then* $\frac{\sqrt{n}}{v_n}(\xi_n - \mathbb{E}(Y)) \xrightarrow[n\to+\infty]{law} \mathcal{N}(0, 1)$.

When $x_\infty = x_\star$, which is nonetheless not required, the limiting variance is optimal in the sense that a crude Monte Carlo computation with the optimal parameter $x_\star$ would have led to the same limiting variance.

2.3.3 A special case: importance sampling for normal random vectors

2.3.3.1 Theoretical framework

We end Section 2.3 by an example of a parametric framework as described in (2.8). We reviewed several ways of building such parametrisations in [L-11] but we chose to focus only on the importance sampling for normal random vectors as it will be further investigated in the next chapter.

Let G be a $d-$dimensional standard normal random vector. For any measurable function $h : \mathbb{R}^d \longrightarrow \mathbb{R}$ such that $\mathbb{E}(|h(G)|) < \infty$,

$$\mathbb{E}\left[h(G)\right] = \mathbb{E}\left[e^{-x\cdot G - \frac{|x|^2}{2}} h(G + x)\right] \quad \text{for all } x \in \mathbb{R}^d. \tag{2.12}$$

Assume we want to compute $\mathbb{E}[f(G)]$ for a measurable function $f : \mathbb{R}^d \longrightarrow \mathbb{R}$ such that $f(G)$ is integrable. By applying equality (2.12) to $h = f$ and $h(u) = f^2(u)\,\mathrm{e}^{-x\cdot u + \frac{|x|^2}{2}}$, one obtains that the variance of the random variable $f(G + x)\,\mathrm{e}^{-x\cdot G - \frac{|x|^2}{2}}$ is given by

$$v(x) = \mathbb{E}\left[f^2(G)\,\mathrm{e}^{-x\cdot G + \frac{|x|^2}{2}}\right] - \mathbb{E}[f(G)]^2.$$

The strict convexity of the function v is already known from [86] for instance, but we can prove a slightly improved version of this result.

Proposition 2.3.8 *Assume that*

$$\mathbb{E}[f(G)] > 0, \tag{2.13}$$

$$\exists \varepsilon > 0,\ \mathbb{E}[|f(G)|^{2+\varepsilon}] < \infty \tag{2.14}$$

Then, v is infinitely continuously differentiable and strongly convex.

2.3.3.2 Bespoke estimators of the optimal variance parameter

We know from Proposition 2.3.8 that the optimal variance parameter is uniquely characterised by $\nabla v(x_\star) = 0$ with

$$\nabla v(x) = \mathbb{E}\left[(x - G) f(G)^2 \, \mathrm{e}^{-x\cdot G + \frac{|x|^2}{2}}\right]. \tag{2.15}$$

If we apply (2.12) again, we obtain an other expression for

$$\nabla v(x) = \mathbb{E}\left[-G f(G + x)^2 \, \mathrm{e}^{-2x\cdot G + |x|^2}\right]. \tag{2.16}$$

Let us introduce the following two functions

$$U^1(x, G) = (x - G) f(G)^2 e^{-x\cdot G + \frac{|x|^2}{2}},$$
$$U^2(x, G) = -G f(G + x)^2 e^{-2x\cdot G + |x|^2}.$$

Using either (2.15) or (2.16), we can write $\nabla v(x) = \mathbb{E}(U^2(x, G)) = \mathbb{E}(U^1(x, G))$ and these two functions U^1 and U^2 fit in the framework of Sections 2.2.3 and 2.3.1.1 and enable us to construct two estimators of $x_\star$ $(X^1_n)_n$ and $(X^2_n)_n$ following (2.4)

$$X^1_{n+1} = \mathcal{T}_{K_{\alpha_n}}\left(X^1_n - \gamma_{n+1} U^1(X^1_n, G_{n+1})\right),$$
$$X^2_{n+1} = \mathcal{T}_{K_{\alpha_n}}\left(X^1_n - \gamma_{n+1} U^2(X^2_n, G_{n+1})\right),$$

where G_n is an i.i.d sequence of random variables with the law of G. We also consider the averaged versions $(\tilde{X}^1_n)_n$ and $(\tilde{X}^2_n)_n$ of these sequences defined as in (2.7). Based on (2.12), we define

$$H(x, G) = f(G + x) e^{-x\cdot G - \frac{|x|^2}{2}}.$$

Corresponding to the different estimators of $x_\star$ listed above, we can define as many approximations of $\mathbb{E}(f(G))$ following (2.10)

$$\xi^1_n = \frac{1}{n}\sum_{i=1}^n H(X^1_{i-1}, G_i), \quad \xi^2_n = \frac{1}{n}\sum_{i=1}^n H(X^2_{i-1}, G_i)$$
$$\tilde{\xi}^1_n = \frac{1}{n}\sum_{i=1}^n H(\tilde{X}^1_{i-1}, G_i), \quad \tilde{\xi}^2_n = \frac{1}{n}\sum_{i=1}^n H(\tilde{X}^2_{i-1}, G_i)$$

where the sequence $(G_i)_i$ has already been used to build the estimators $(X^1_n)_n$, $(X^2_n)_n$, $(\tilde{X}^1_n)_n$ and $(\tilde{X}^2_n)_n$. From Proposition 2.3.8 and Corollary 2.2.3 and Theorems 2.3.3 and 2.3.5, we can deduce the following result.

Theorem 2.3.9 *If there exists $\varepsilon > 0$ such that $\mathbb{E}[f(G)^{4+\varepsilon}] < \infty$ then, the sequences $(X^1_n)_n$, $(X^2_n)_n$, $(\tilde{X}^1_n)_n$ and $(\tilde{X}^2_n)_n$ converge a.s. to $x_\star$ for any increasing sequence of compact sets $(K_j)_j$ satisfying (2.3) and the adaptive estimator $(\xi^1_n)_n$, $(\xi^2_n)_n$, $(\tilde{\xi}^1_n)_n$ and $(\tilde{\xi}^2_n)_n$ converge to $\mathbb{E}[f(G)]$ and are asymptotically normal with optimal limiting variance $v(x_\star)$.*

2.3.3.3 Complexity of the different approximations

From Theorem 2.3.9, we know that the adaptive estimators $(\xi_n^1)_n, (\xi_n^2)_n, (\tilde{\xi}_n^1)_n, (\tilde{\xi}_n^2)_n$ all converge with the same rate $\sqrt{v(x_\star)/n}$ but they do not have the same computational cost. First, let us concentrate on $(\xi_n^1)_n$ and $(\tilde{\xi}_n^1)_n$. At each iteration i, the function f has to be computed twice : once at the point $G_{i+1} + X_i^1$ (or $G_{i+1} + \tilde{X}_i^1$) to update the Monte Carlo estimator and once at the point G_{i+1} to update X_{i+1}^1. Hence, the computation of ξ_n^1 or $\tilde{\xi}_n^1$ requires $2n$ evaluations of the function f. Similarly, the computation of $\tilde{\xi}_n^2$ requires $2n$ evaluations of the function f. Closely looking at the computation of $(\xi_n^2)_n$ immediately highlights the benefit of having put the parameter x back into the function f in the expression of ∇v : the updates of ξ_{i+1}^2 and X_{i+1}^2 both use the same evaluation of the function f. Hence, the computation of ξ_n^2 only needs n evaluations of the function f instead of $2n$ for all the other algorithms. Obviously, the computational costs of the different estimators cannot really be reduced to the number of times the function f is evaluated so one should not expect that computing ξ_n^2 is twice less costly than the other estimators but we will see in the examples below that the estimator ξ_n^2 is indeed faster than the others.

To shortly conclude on the complexity of the different algorithms, be they sequential or adaptive, one should bear in mind that all the estimators except $(\xi_n^2)_n$ roughly require twice the computational time of the crude Monte-Carlo method.

2.3.3.4 Application to the pricing of basket options in a local volatility model

Now, we compare the different algorithms on multi-asset options. The quantity "Var MC" denotes the variance of the crude Monte Carlo estimator computed on-line on a single run of the algorithm. The variance denoted "Var ξ^2" (resp. "Var $\tilde{\xi}^2$") is the variance of the ADIS algorithm (see Algorithm 2.3.2) which uses $(X_n^2)_n$ (resp. $(\tilde{X}_n^2)_n$) to estimate $x_\star$. These variances are computed using the on-line estimator given by Corollary 2.3.7.

We consider options with payoffs of the form $(\sum_{i=1}^d \omega^i S_T^i - K)_+$ where $(\omega^1, \ldots, \omega^d)$ is a vector of algebraic weights (enabling us to consider exchange options).

ρ	K	γ	Price	Var MC	Var ξ^2	Var $\tilde{\xi}^2$
0.1	45	1	7.21	12.24	1.59	1.10
	55	10	0.56	1.83	0.19	0.14
0.2	50	0.1	3.29	13.53	1.82	1.76
0.5	45	0.1	7.65	43.25	6.25	4.97
	55	0.1	1.90	14.74	1.91	1.4
0.9	45	0.1	8.24	69.47	10.20	7.78
	55	0.1	2.82	30.87	2.7	2.6

Table 2.1: Basket option in dimension $d = 40$ with $r = 0.05$, $T = 1$, $S_0^i = 50$, $\sigma^i = 0.2$, $\omega^i = \frac{1}{d}$ for all $i = 1, \ldots, d$ and $n = 100\,000$.

Estimators	MC	ξ^2	$\tilde{\xi}^2$
CPU time	0.85	0.9	1.64

Table 2.2: CPU times for the option of Table 2.1.

The results of Table 2.1 indicate that the adaptive algorithm using an averaging stochastic approximation outperforms not only the crude Monte Carlo approach but also the adaptive algorithms using

non-averaging stochastic approximation. Nonetheless, these good results in terms of variance reduction must be considered together with their computation costs reported in Table 2.2. As explained in Section 2.3.3.3, we notice that the computational cost of the estimator ξ^2 is very close to the one of the crude Monte Carlo estimator because the implementation made the most of the fact that the updates of ξ_{i+1}^2 and X_{i+1}^2 both need to evaluate the function f at the same point. Since this implementation trick cannot be applied to $\tilde{\xi}^2$, the adaptive algorithm using an averaging stochastic approximation is twice slower. For a given precision, the adaptive algorithm is between 5 and 10 times faster.

2.4 Conclusion

In this chapter, we have presented theoretical results on the convergence of stochastic approximation with random truncations to handle fast growing functions. These algorithms aim at computing the minimum of strictly convex and continuously differentiable functions given as expectations, which typically come up when using importance sampling. In the second half of the chapter, we have explained how to use stochastic approximation to implement adaptive Monte Carlo methods based on importance sampling and we have specially focused on the normal random vector framework. The method has proved to be efficient but requires fine tuning of the gain sequence $(\gamma_n)_n$. In the next chapter, we will propose a more robust algorithm for computing the optimal importance sampling parameter.

Chapter 3

The *Sample Average Approximation* approach to Importance Sampling

In this chapter, we develop a parametric approach to adaptive importance sampling, which uses *Sample Average Approximation* (see [83]) to compute the optimal importance sampling parameter. We focus more specifically on two cases: the Gaussian vectors (see [L-12]) and jump diffusion processes (see [L-5]). The algorithm developed for the Gaussian framework is currently being used by *Natixis* for their foreign exchange derivative pricer and *Mentor Graphics* for rare event simulation when designing electronic circuits.

3.1 The importance sampling framework

3.1.1 A parametric approach

Let E be a Polish space and ν a measure defined on E. We consider the problem of computing $\mathbb{E}[\varphi(X)]$ where the random variable X taking values in E has a density function $g : E \longmapsto \mathbb{R}_+$ with respect to the measure ν. Consider a family of random variables $(X^{(\theta)})_{\theta\in\mathbb{R}^p}$ with values in E and density functions $(f_\theta)_{\theta\in\mathbb{R}^p}$ with respect to ν satisfying $\mathrm{supp}(g) \subset \mathrm{supp}(f_\theta)$ for all $\theta \in \mathbb{R}^p$. Then, we can write

$$\mathbb{E}[\varphi(X)] = \int_E \varphi(x)g(x)d\nu(x) = \int_E \varphi(x)\frac{g(x)}{f_\theta(x)}f_\theta(x)d\nu(x)$$
$$\mathbb{E}[\varphi(X)] = \mathbb{E}\left[\varphi(X^{(\theta)})\frac{g(X^{(\theta)})}{f_\theta(X^{(\theta)})}\right]. \tag{3.1}$$

From a practical point of view, one has to find the best density function h in order to maximize the accuracy of the Monte Carlo estimator. First, we need to make precise how to measure the quality of a density. Two criteria are usually considered to define the best density function: the variance or the cross–entropy.

The variance. As the convergence of the Monte Carlo estimator is governed by the central limit theorem, it is quite natural to try to find the density function f_θ minimizing the variance of the estimator

$$v(\theta) = \mathrm{Var}\left(\varphi(X^{(\theta)})\frac{g(X^{(\theta)})}{f_\theta(X^{(\theta)})}\right) = \mathbb{E}\left[\varphi(X^{(\theta)})^2\left(\frac{g(X^{(\theta)})}{f_\theta(X^{(\theta)})}\right)^2\right] - \left(\mathbb{E}[\varphi(X)]\right)^2.$$

Only the second moment depends on the density function f_θ and moreover using (3.1) again, we obtain

$$v(\theta) = \mathbb{E}\left[\varphi(X)^2 \frac{g(X)}{f_\theta(X)}\right] - \left(\mathbb{E}[\varphi(X)]\right)^2. \tag{3.2}$$

With this expression, we managed to decouple the objective function φ and the importance sampling density function f_θ. The second moment naturally writes as an expectation w.r.t the density function g and the auxiliary density function f_θ is only involved in the importance sampling weight. The decoupling between φ and f_θ will play a crucial role to prove regularity properties for the function v.

From (3.2), we easily deduce that the optimal change of measure minimizing

$$\mathbb{E}\left[\varphi(X)^2 \frac{g(X)}{h(X)}\right]$$

over all possible functions h is defined by

$$h^\star(x) = \frac{\varphi(x)g(x)}{\mathbb{E}[\varphi(X)]}. \tag{3.3}$$

This choice actually leads to a zero variance estimator but cannot be used in practice as it involves the expectation to be computed as a scaling factor. If the optimization problem $\min_\theta \mathbb{E}\left[\varphi(X)^2 \frac{g(X)}{f_\theta(X)}\right]$ is well–posed (strongly convex for instance), we can solve it directly and efficiently. When this approach reveals too difficult, the cross entropy method (see de Boer et al. [32]) can be used.

The cross entropy. It amounts to minimizing the distance between $h^\star$ and the set of functions $(f_\theta)_\theta$ defined by

$$\mathcal{D}(h^*, f_\theta) = \int_E \log \frac{h^\star(x)}{f_\theta(x)} h^*(x) d\nu(x) = \int_E h^\star(x) \log h(x) d\nu(x) - \int_E h^*(x) \log f(x) d\nu(x).$$

Minimizing $\mathcal{D}(h^\star, f_\theta))$ over θ boils down to solving

$$\max_\theta \int_e h^\star(x) \log f_\theta(x) d\nu(x).$$

Substituting $h^\star$ with its expression from (3.3), we obtain the equivalent maximization problem

$$\max_\theta \int_E \varphi(x)g(x) \log f_\theta(x) d\nu(x) = \max_\theta \mathbb{E}[\varphi(X) \log f_\theta(X)].$$

Remark. The cross entropy approach is generally better suited for rare event simulation than the variance criterion. However, these two approaches have been extensively tested in the Gaussian framework for both variance reduction and rare event simulation in collaboration with *Mentor Graphics* and we came to the conclusion that the variance criterion was more efficient. It leads to a far more robust algorithm (see Section 3.3.3), which converges in very few gradient steps (usually 5 steps are enough to claim convergence).

3.1.2 The exponential change of measure

A very common way of building a set of parametrized densities is to use the Esscher transform leading to the exponential change of measure. Let $\psi(\theta) = \log \mathbb{E}[\mathrm{e}^{\theta \cdot X}]$ be the cumulative generating function of X. We assume that $\psi(\theta) < \infty$ for all $\theta \in \mathbb{R}^d$, which implies that ψ is infinitely differentiable. Using Hölder's inequality, one can prove that ψ is convex. We define the family $(f_\theta)_\theta$ by

$$f_\theta(x) = g(x)\, \mathrm{e}^{\theta \cdot x - \psi(x)}, \quad x \in \mathbb{R}^d.$$

From (3.1) and (3.2), we deduce that

$$\begin{aligned} \mathbb{E}[\varphi(X)] &= \mathbb{E}\left[\varphi(X^{(\theta)})\, \mathrm{e}^{-\theta \cdot X^{(\theta)} + \psi(\theta)}\right] \\ v(\theta) &= \mathbb{E}[\varphi(X)^2\, \mathrm{e}^{-\theta \cdot X + \psi(\theta)}] - \mathbb{E}[f(X)]^2. \end{aligned}$$

Proposition 3.1.1 *Assume that*

$$\exists \gamma > 0,\ \mathbb{E}[|\varphi(X)|^{2+\gamma}] < \infty; \tag{3.4}$$

$$\lim_{|\theta| \to \infty} f_\theta(x) = 0 \quad \textit{for all } x \in \mathbb{R}^d. \tag{3.5}$$

Then, v is infinitely differentiable, convex, $\lim_{|\theta| \to \infty} v(\theta) = \infty$ *and*

$$\begin{aligned} \nabla v(\theta) &= \mathbb{E}\left[(\nabla \psi(\theta) - X)\varphi(X)^2\, \mathrm{e}^{-\theta \cdot X + \psi(\theta)}\right] \\ \nabla^2 v(\theta) &= \mathbb{E}\left[(\nabla^2 \psi(\theta) + (\nabla \psi(\theta) - X)(\nabla \psi(\theta) - X)^T)\varphi(X)^2\, \mathrm{e}^{-\theta \cdot X + \psi(\theta)}\right]. \end{aligned}$$

From Proposition 3.1.1,

$$\nabla^2 v(\theta) = \mathbb{E}\left[\nabla^2 \psi(\theta)\varphi(X)^2\, \mathrm{e}^{-\theta \cdot X + \psi(\theta)}\right] + \mathbb{E}\left[(\nabla \psi(\theta) - X)(\nabla \psi(\theta) - X)^T \varphi(X)^2\, \mathrm{e}^{-\theta \cdot X + \psi(\theta)}\right].$$

The second term on the r.h.s is a positive semi–definite matrix. Assume there exists $\delta > 0$ such that $\theta \mapsto \psi(\theta) - \frac{\delta}{2}|\theta|^2$ is convex, then

$$\nabla^2 v(\theta) \geq \delta \mathbb{E}\left[\varphi(X)^2\, \mathrm{e}^{-\theta \cdot X + \psi(\theta)}\right] I_d \geq \delta \mathbb{E}[|\varphi(X)|]^2 I_d$$

where the last part ensues from the Cauchy Schwartz inequality. Hence, v is strongly convex as soon as ψ is strongly convex and $\mathbb{P}(\varphi(X) \neq 0) > 0$.

Now, we focus on two specific examples, which will be further investigated in the coming sections.

The Gaussian distribution. Let X be a standard normal random vector with values in $\mathbb{R}^d$ and choose for ν the Lebesgue measure on $\mathbb{R}^d$. Then, $g(x) = (2\pi)^{-\frac{1}{2d}}\, \mathrm{e}^{-|x|^2/2}$ and $\psi(\theta) = \frac{|\theta|^2}{2}$ for $\theta \in \mathbb{R}^d$. The family $(f_\theta)_\theta$ is defined by

$$f_\theta(x) = (2\pi)^{-\frac{1}{2d}}\, \mathrm{e}^{-|x|^2/2}\, \mathrm{e}^{\theta \cdot x - |\theta|^2/2} = (2\pi)^{-\frac{1}{2d}}\, \mathrm{e}^{|x-\theta|^2/2},$$

which corresponds to the density function of normal random vector with mean θ and identity covariance matrix. In the Gaussian case, the Esscher transform actually coincides with the well–known mean shift approach.

The Poisson distribution. Let X be a Poisson distribution with parameter $\lambda > 0$ and choose for ν the counting measure on $\mathbb{N}$. Then, $g(x) = \frac{\lambda^x \, \mathrm{e}^{-\lambda}}{x!}$ for $x \in \mathbb{N}$ and $\psi(\theta) = \lambda(\mathrm{e}^{\theta} - 1)$ for $\theta \in \mathbb{R}$. The family $(f_\theta)_\theta$ is defined by

$$f_\theta(x) = \frac{\lambda^x \, \mathrm{e}^{-\lambda}}{x!} \, \mathrm{e}^{\theta x - \lambda(\mathrm{e}^{\theta} - 1)}, \quad \text{for } x \in \mathbb{N}.$$

Take $\theta = \log \frac{\mu}{\lambda}$ for $\mu > 0$, then

$$f_\theta(x) = \frac{\mu^x \, \mathrm{e}^{-\mu}}{x!},$$

which corresponds to the probability mass function of a Poisson random variable with parameter μ. This example can naturally be extended to the case of a random vector with independent Poisson components.

3.2 Importance sampling for the mixed Gaussian Poisson framework

In the rest of this chapter, we focus on the specific Gaussian Poisson framework. Les G be a standard Gaussian vector in $\mathbb{R}^d$ and $N^\mu = (N_1^{\mu_1}, \cdots, N_p^{\mu_p})$ a vector of p independent Poisson random variables with parameters $\mu = (\mu_1, \ldots, \mu_p)$. The random variable N^m will be called a Poisson random vector. Moreover, we assume that G and N^μ are independent. The motivation for studying such a framework comes from the discretization of a jump diffusion process. In this case, the size $p + d$ of the problem can become very large as the effective dimension is the number of time steps times the dimension of the driving Brownian and Poisson process. We focus on the computation of

$$\mathcal{E} = \mathbb{E}[f(G, N^\mu)] \tag{3.6}$$

where $f : \mathbb{R}^d \times \mathbb{N}^p \longrightarrow \mathbb{R}$ satisfies $\mathbb{E}[|f(G, N^\mu)|] < \infty$. Based on the exponential change of measure, we deduce the following importance sampling formula.

Lemma 3.2.1 *For any measurable function $h : \mathbb{R}^d \times \mathbb{N}^p \longrightarrow \mathbb{R}$ either positive or such that $\mathbb{E}[|h(G, N^\mu)|] < \infty$, one has for all $\theta \in \mathbb{R}^d, \lambda \in (]0, +\infty[)^p$*

$$\mathbb{E}[h(G, N^\mu)] = \mathbb{E}\left[h(G + \theta, N^\lambda) \, \mathrm{e}^{-\theta \cdot G - \frac{|\theta|^2}{2}} \prod_{i=1}^{p} \mathrm{e}^{\lambda_i - \mu_i} \left(\frac{\mu_i}{\lambda_i} \right)^{N_i^{\lambda_i}} \right] \tag{3.7}$$

where N^λ is a Poisson random vector with parameter $\lambda = (\lambda_1, \ldots, \lambda_p)$.

When the expectation $\mathcal{E}$ is computed using a Monte Carlo method, the Central Limit Theorem advises to use the representation of $f(G, N^\mu)$ with the smallest possible variance which is achieved by choosing the parameters (θ, λ) which minimize the variance of of $f(G + \theta, N^\lambda) \, \mathrm{e}^{-\theta \cdot G - \frac{|\theta|^2}{2}} \prod_{i=1}^{p} \mathrm{e}^{\lambda_i - \mu_i} \left(\frac{\mu_i}{\lambda_i} \right)^{N_i^{\lambda_i}}$. This raises several questions which are investigated in the chapter. Does the variance of $f(G + \theta, N^\lambda) \, \mathrm{e}^{-\theta \cdot G - \frac{|\theta|^2}{2}} \prod_{i=1}^{p} \mathrm{e}^{\lambda_i - \mu_i} \left(\frac{\mu_i}{\lambda_i} \right)^{N_i^{\lambda_i}}$ admits a unique minimizer? If so, how can it be computed numerically and how to make the most of it in view of a further Monte Carlo computation?

These questions are quite natural in the context of Monte Carlo computations and have already been widely discussed in the pure Gaussian framework. The first applications to option pricing of some adaptive Monte Carlo methods based on importance sampling goes back to the papers of [3, 4]. These papers were based on a change of mean for the Gaussian random normal vectors and the optimal parameter was searched for by using some stochastic approximation algorithm with random truncations. This approach was investigated in Section 2.3. To circumvent the delicate tuning of stochastic approximation, we propose to use sample average approximation instead, which basically relies on deterministic optimization techniques. An alternative to random truncations was studied by [70] who managed to modify the initial problem in order ta apply the more standard Robbins Monro algorithm. Not only have they applied this to the Gaussian framework but they have considered a few examples of Levy processes relying on the Esscher transform to introduce a free parameter. The idea of using the Esscher transform was also extensively investigated by [56, 57, 58].

3.2.1 Computing the optimal importance sampling parameters

Thanks to Lemma 3.2.1, the expectation $\mathcal{E}$ can be written

$$\mathcal{E} = \mathbb{E}\left[f(G+\theta, N^\lambda)\, \mathrm{e}^{-\theta\cdot G - \frac{|\theta|^2}{2}} \prod_{i=1}^p \mathrm{e}^{\lambda_i-\mu_i} \left(\frac{\mu_i}{\lambda_i}\right)^{N_i^{\lambda_i}} \right], \quad \forall\, \theta\in\mathbb{R}^d, \lambda \in {\mathbb{R}_+^*}^p.$$

Note that for the particular choice of $\theta = 0$ and $\lambda = \mu$, we recover Equation (3.6).

The convergence rate of a Monte Carlo estimator of $\mathcal{E}$ based on this new representation is governed by the variance of $f(G+\theta, N^\lambda)\, \mathrm{e}^{-\theta\cdot G - \frac{|\theta|^2}{2}} \prod_{i=1}^p \mathrm{e}^{\lambda_i-\mu_i} \left(\frac{\mu_i}{\lambda_i}\right)^{N_i^{\lambda_i}}$ which can be written in the form $v(\theta,\lambda) - \mathcal{E}^2$ where

$$v(\theta,\lambda) = \mathbb{E}\left[f(G, N^\mu)^2\, \mathrm{e}^{-\theta\cdot G + \frac{|\theta|^2}{2}} \prod_{i=1}^p \mathrm{e}^{\lambda_i-\mu_i} \left(\frac{\mu_i}{\lambda_i}\right)^{N_i^{\mu_i}} \right]. \tag{3.8}$$

To keep equations a bit more concise, we introduce the function $F : \mathbb{R}^d\times\mathbb{R}^d\times\mathbb{N}\times{\mathbb{R}_+^*}^p\times{\mathbb{R}_+^*}^p \longrightarrow \mathbb{R}$ defined by

$$F(g,\theta,n,\lambda,\mu) = f(g,n)^2\, \mathrm{e}^{-\theta\cdot g + \frac{|\theta|^2}{2}} \prod_{i=1}^p \mathrm{e}^{\lambda_i-\mu_i} \left(\frac{\mu_i}{\lambda_i}\right)^{n_i}. \tag{3.9}$$

This expression of v is easily obtained by applying Lemma 3.2.1 to the function $h(g,n) = f(g+\theta,n)^2\, \mathrm{e}^{-2\theta\cdot g - |\theta|^2} \prod_{i=1}^p \mathrm{e}^{2(\lambda_i-\mu_i)} \left(\frac{\mu_i}{\lambda_i}\right)^{2n_i}$. Applying the change of measure backward after computing the variance enables us to write the variance in a form which does not involve the parameters θ and λ in the arguments of the function f. From Proposition 3.1.1, we deduce

Proposition 3.2.2 *Assume that*

(H3.1) *i.* $\exists (n_1,\ldots,n_p)\in\mathbb{N}^{*p},\ s.t.\ \mathbb{P}(|f(G,(n_1,\ldots,n_p))|>0)>0$

ii. $\exists\gamma>0, \quad \mathbb{E}\left[|f(G,N^\mu)|^{2+\gamma}\right] < \infty.$

Then, the function v is infinitely continuously differentiable, strongly convex and moreover the gradient vectors are given by

$$\nabla_\theta v(\theta,\lambda) = \mathbb{E}\left[(\theta - G)\, F(G,\theta,N^\mu,\lambda,\mu)\right]; \quad \nabla_\lambda v(\theta,\lambda) = \mathbb{E}\left[a(N^\mu,\lambda) F(G,\theta,N^\mu,\lambda,\mu)\right] \tag{3.10}$$

where the vector $a(N^\mu,\lambda)=\left(1-\frac{N_1^{\mu_1}}{\lambda_1},\dots,1-\frac{N_p^{\mu_p}}{\lambda_p}\right)^T$. *The second derivatives are defined by*

$$\nabla^2_{\theta,\theta}v(\theta,\lambda) =\mathbb{E}\left[\left(I_d+(\theta-G)(\theta-G)^T\right)F(G,\theta,N^\mu,\lambda,\mu)\right], \tag{3.11}$$

$$\nabla^2_{\theta,\lambda}v(\theta,\lambda) =\mathbb{E}\left[(\theta-G)a(N^\mu,\lambda)^T F(G,\theta,N^\mu,\lambda,\mu)\right], \tag{3.12}$$

$$\nabla^2_{\lambda,\lambda}v(\theta,\lambda) =\mathbb{E}\left[\left(D+a(N^\mu,\lambda)a(N^\mu,\lambda)^T\right)F(G,\theta,N^\mu,\lambda,\mu)\right] \tag{3.13}$$

where the diagonal matrix D *is defined by* $D=\mathrm{diag}_p\left(\frac{N_1^{\mu_1}}{\lambda_1^2},\dots,\frac{N_p^{\mu_p}}{\lambda_p^2}\right)$.

As a consequence, the function v admits a unique minimizer $(\theta_\star,\lambda_\star)$ defined by $\nabla_\theta v(\theta_\star,\lambda_\star)=\nabla_\lambda v(\theta_\star,\lambda_\star)=0$. The characterization of $(\theta_\star,\lambda_\star)$ as the unique minimizer of a strongly convex function is very appealing but there is no hope to compute the gradient of v in a closed form, so we will need to resort to some kind of approximations before running the optimization step. In many situations, such as the discretization of a jump diffusion, it is advisable to reduce the dimension of the space in which the optimization problem is solved. We will see in the examples that it leads to huge computational time savings and do not deteriorate the optimal variance significantly.

Reducing the dimension of the optimization problem. Let $0<d'\leq d$ and $0<p'\leq p$ be the reduced dimensions. Instead of searching for the best importance sampling parameter (θ,λ) in the whole space $\mathbb{R}^d\times{\mathbb{R}_+^*}^p$, we consider the subspace $\{(A\vartheta,B\lambda)\ :\ \vartheta\in\mathbb{R}^{d'},\ \lambda\in{\mathbb{R}_+^*}^{p'}\}$ where $A\in\mathbb{R}^{d\times d'}$ is a matrix with rank $d'\leq d$ and $B\in{\mathbb{R}_+^*}^{p\times p'}$ a matrix with rank $p'\leq p$. Since all the coefficients of B are non negative, for all $\vartheta\in{\mathbb{R}_+^*}^{p'}$, $B\vartheta\in{\mathbb{R}_+^*}^{p}$.

For such matrices A and B, we introduce the function $v^{A,B}:\mathbb{R}^{d'}\times{\mathbb{R}_+^*}^{p'}\longmapsto\mathbb{R}$ defined by

$$v^{A,B}(\vartheta,\lambda)=v(A\vartheta,B\lambda). \tag{3.14}$$

The function $v^{A,B}$ inherits from the regularity and convexity properties of v. Hence, from Proposition 3.2.2, we know that $v^{A,B}$ is continuously infinitely differentiable and strongly convex. As a consequence, there exists a unique couple of minimizers $(\vartheta_\star^{A,b},\lambda_\star^{A,B})$ such that $v^{A,B}(\vartheta_\star^{A,B},\lambda_\star^{A,B})=\inf_{\vartheta\in\mathbb{R}^{d'},\lambda\in{\mathbb{R}_+^*}^{p'}}v^{A,B}(\vartheta,\lambda)$. For the particular choices $A=I_d$, $B=I_p$, $d=d'$ and $p=p'$, the functions v^{I_d,I_p} and v coincide.

The Esscher transform as a way to reduce the dimension. Consider a two dimensional process $(X_t)_{t\leq T}$ of the form $X_t=(W_t,\tilde{N}_t^{\tilde{\mu}})$ where W is a real Brownian motion and $\tilde{N}^{\tilde{\mu}}$ is a Poisson process with intensity $\tilde{\mu}$. The Esscher transform applied to X yields that for any nonnegative function h, we have the following equality $\forall\ \alpha\in\mathbb{R},\tilde{\lambda}\in\mathbb{R}_+^*$,

$$\mathbb{E}[h((W_t,\tilde{N}_t^{\tilde{\mu}})_{t\leq T})]=\mathbb{E}\left[h((W_t+\alpha t,\tilde{N}^{\tilde{\lambda}})_{t\leq T})\,\mathrm{e}^{-\alpha W_T-\frac{|\alpha|^2T}{2}}\,\mathrm{e}^{T(\tilde{\lambda}-\tilde{\mu})}\left(\frac{\tilde{\mu}}{\tilde{\lambda}}\right)^{\tilde{N}_T^{\tilde{\lambda}}}\right]$$

Let $0=t_0<\cdots<t_p=T$ be a time grid of $[0,T]$. If we consider the vector G (resp. N^μ) as the increments of W (resp. $\tilde{N}^{\tilde{\mu}}$) on the grid, we can recover a particular form of Equation (3.7) with $A,B\in\mathbb{R}^p$ given by

$$A=\left(\sqrt{t_1},\sqrt{t_2-t_1},\dots,\sqrt{t_p-t_{p-1}}\right)^T;\quad B=(t_1,t_2-t_1,\dots,t_p-t_{p-1})^T.$$

3.2.2 Tracking the optimal importance sampling parameter

The optimal importance sampling parameter $(\theta_\star\lambda_\star)$ are characterized as the unique zero of an expectation, which is the typical framework for applying stochastic approximation, which was the point of view adopted in Chapter 2. In this work, we adopt a totally different point of view often called *sample average approximation*, which basically consists in first replacing expectations by sample averages and then using deterministic optimization techniques on these empirical means.

Let $(G^j)_{j\geq 1}$ be a sequence of $d-$dimensional independent and identically distributed standard normal random variables. We also introduce $(N^{\mu,j})_{j\geq 1}$ a sequence of $p-$dimensional independent and identically distributed Poisson random vectors with intensity μ. For $m \geq 1$, we introduce the sample average approximation of the function $v^{A,B}$ defined by

$$v_m^{A,B}(\vartheta,\lambda) = \frac{1}{m}\sum_{j=1}^{m} f(G^j, N^{\mu,j})^2\, \mathrm{e}^{-A\vartheta\cdot G^j + \frac{|A\vartheta|^2}{2}} \prod_{i=1}^{p} \mathrm{e}^{(B\lambda)_i-\mu_i} \left(\frac{\mu_i}{(B\lambda)_i}\right)^{N_i^{\mu_i,j}}. \tag{3.15}$$

For n large enough, $f(G^j, N^{\mu,j}) \neq 0$ for some index $j \in \{1,\ldots,m\}$ and the approximation $v_n^{A,B}$ is also strongly convex and hence admits a unique minimizer $(\vartheta_m^{A,B}, \lambda_m^{A,B})$ defined by

$$(\vartheta_m^{A,B}, \lambda_m^{A,B}) = \operatorname*{arginf}_{\vartheta\in\mathbb{R}^{d'},\lambda\in{\mathbb{R}_+^*}^{p'}} v_m^{A,B}(\vartheta,\lambda). \tag{3.16}$$

Proposition 3.2.3 *Under Assumption ($\mathcal{H}$3.1), the sequence of random functions $(v_n^{A,B})_n$ converges a.s. locally uniformly to the continuous function $v^{A,B}$.*

To prove this result, we use the uniform strong law of large numbers recalled hereafter, see for instance [83, Lemma A1]. This result is also a consequence of the strong law of large numbers in Banach spaces [69, Corollary 7.10, page 189].

Lemma 3.2.4 *Let $(X_i)_{i\geq 1}$ be a sequence of i.i.d. $\mathbb{R}^m$-valued random vectors, E an open set of $\mathbb{R}^d$ and $h : E\times\mathbb{R}^m \to \mathbb{R}$ be a measurable function. Assume that*

- *a.s., $\chi \in E \mapsto h(\chi, X_1)$ is continuous,*
- *for all compact sets K of $\mathbb{R}^d$ such that $K \subset E$, $\mathbb{E}\left(\sup_{\chi\in K}|h(\chi,X_1)|\right) < +\infty$.*

Then, a.s. the sequence of random functions $\chi \in K \mapsto \frac{1}{n}\sum_{i=1}^n h(\chi, X_i)$ converges locally uniformly to the continuous function $\chi \in E \mapsto \mathbb{E}(h(\chi, X_1))$.

Proof (Proof of Proposition 3.2.3). It is sufficient to prove the result for v_n and it will hold for $v_n^{A,B}$. Let $M > \underline{m} > 0$. For all (θ,λ) such that $|(\theta,\lambda)| \leq M$ and $d_0(\lambda) > \underline{m}$, we have

$$\mathrm{e}^{-\theta\cdot G+\frac{|\theta|^2}{2}} \prod_{i=1}^{p} \mathrm{e}^{\lambda_i-\mu_i}\left(\frac{\mu_i}{\lambda_i}\right)^{N_i^{\mu_i}} \leq \prod_{k=1}^{d}(\mathrm{e}^{-MG_k}+\mathrm{e}^{MG_k})\,\mathrm{e}^{\frac{M^2}{2}} \prod_{i=1}^{p}\mathrm{e}^{M-\mu_i}\left(\frac{\mu_i}{\underline{m}}\right)^{N_i^{\mu_i}}.$$

The r.h.s. is integrable by ($\mathcal{H}$3.1) and Hölder's inequality; hence, we can apply Lemma 3.2.4. ■

Proposition 3.2.5 *Under Assumption ($\mathcal{H}$3.1), the pair $(\vartheta_m^{A,B}, \lambda_m^{A,B})$ converges a.s. to $(\vartheta_\star^{A,B}, \lambda_\star^{A,B})$ as $m \to +\infty$. Moreover, if*

($\mathcal{H}$3.2) $\exists\delta > 0, \quad \mathbb{E}\left[|f(G,N^\mu)|^{4+\delta}\right] < \infty,$

$\sqrt{m}\left((\vartheta_m^{A,B},\lambda_m^{A,B})-(\vartheta_\star^{A,B},\lambda_\star^{A,B})\right)$ *converges in law to the normal distribution* $\mathcal{N}_{d+p}(0,\Gamma)$ *where*

$$\Gamma=\left(\nabla^2 v^{A,B}(\vartheta_\star^{A,B},\lambda_\star^{A,B})\right)^{-1}\mathrm{Cov}(\nabla F(G,A\vartheta_\star^{A,B},N^\mu,B\lambda_\star^{A,B}))\left(\nabla^2 v^{A,B}(\vartheta_\star^{A,B},\lambda_\star^{A,B})\right)^{-1}.$$

Condition ($\mathcal{H}3.2$) ensures that the covariance matrix $\mathrm{Cov}(\nabla F(G,A\vartheta_\star^{A,B},N^\mu,B\lambda_\star^{A,B}))$ is well defined. The non singularity of the matrix $\nabla^2 v^{A,B}(\vartheta_\star^{A,B},\lambda_\star^{A,B})$ is guaranteed by the strict convexity of v. By combining Propositions 3.2.3 and 3.2.5, we can state the following result

Corollary 3.2.6 *Under Assumption ($\mathcal{H}3.1$),* $v_m^{A,B}(\vartheta_m^{A,B},\lambda_m^{A,B})$ *converge a.s. to* $v^{A,B}(\vartheta_\star^{A,B},\lambda_\star^{A,B})$ *as* $n\to+\infty$.

The second part of Proposition 3.2.5 is a consequence of [83, Theorem A2]. Hence, we only prove the first a.s. convergence. If the sequence $(\vartheta_m^{A,B},\lambda_m^{A,B})_m$ were obtained as the solution of a sequence of optimization problems solved under the compactness constraint, the a.s. convergence would follow from [83, Theorem A1] as a direct consequence of the uniform strong law of large numbers.

Proof (Proof of Proposition 3.2.5). Let $\varepsilon>0$. We define a compact neighbourhood $\mathcal{V}_\varepsilon$ of $(\vartheta_\star,\lambda_\star)$

$$\mathcal{V}_\varepsilon\stackrel{\Delta}{=}\left\{(\vartheta,\lambda)\in\mathbb{R}^{d'}\times\mathbb{R}^{p'}\ :\ |(\vartheta,\lambda)-(\vartheta_\star,\lambda_\star)|\le\varepsilon\right\}. \tag{3.17}$$

In the following, we assume that ε is small enough, so that $\mathcal{V}_\varepsilon$ is included in $\mathbb{R}^{d'}\times{\mathbb{R}_+^*}^{p'}$.

By the strict convexity and the continuity of $v^{A,B}$,

$$\alpha\stackrel{\Delta}{=}\inf_{(\vartheta,\lambda)\in\mathcal{V}_\varepsilon^c} v^{A,B}(\vartheta,\lambda)-v^{A,B}(\vartheta_\star^{A,B},\lambda_\star^{A,B})>0.$$

The local uniform convergence of $v_m^{A,B}$ to $v^{A,B}$ ensures that for some n_α sufficiently large,

$$\forall m\ge m_\alpha,\ \forall(\vartheta,\lambda)\in\mathcal{V}_\varepsilon,\ |v_m^{A,B}(\vartheta,\lambda)-v^{A,B}(\vartheta,\lambda)|\le\frac{\alpha}{3}. \tag{3.18}$$

For $m\ge m_\alpha$ and $(\vartheta,\lambda)\notin\mathcal{V}_\varepsilon$, we define $(\vartheta_\varepsilon^{A,B},\lambda_\varepsilon^{A,B})\in\mathcal{V}_\varepsilon$ by

$$(\vartheta_\varepsilon^{A,B},\lambda_\varepsilon^{A,B})\stackrel{\Delta}{=}\left(\vartheta_\star^{A,B}+\varepsilon\frac{\vartheta-\vartheta_\star^{A,B}}{|(\vartheta-\vartheta_\star^{A,B},\lambda-\lambda_\star^{A,B})|},\lambda_\star^{A,B}+\varepsilon\frac{\mu-\lambda_\star^{A,B}}{|(\vartheta-\vartheta_\star^{A,B},\lambda-\lambda_\star^{A,B})|}\right).$$

Using the convexity of $v_m^{A,B}$ for the first inequality and Equation (3.18) for the second one, we deduce

$$\begin{aligned}v_m^{A,B}(\vartheta,\lambda)-v_m^{A,B}(\vartheta_\star^{A,B},\lambda_\star^{A,B})&\ge\frac{|(\vartheta-\vartheta_\star^{A,B},\lambda-\lambda_\star^{A,B})|}{\varepsilon}\left[v_m^{A,B}(\vartheta_\varepsilon^{A,B},\lambda_\varepsilon^{A,B})-v_m^{A,B}(\vartheta_\star^{A,B},\lambda_\star^{A,B})\right]\\&\ge\left[v^{A,B}(\vartheta_\varepsilon^{A,B},\lambda_\varepsilon^{A,B})-v^{A,B}(\vartheta_\star^{A,B},\lambda_\star^{A,B})-\frac{2\alpha}{3}\right]\ge\frac{\alpha}{3}.\end{aligned}$$

The optimality of $(\vartheta_m^{A,G},\lambda_m^{A,B})$ yields that $v_m^{A,B}(\vartheta_m^{A,B},\lambda_m^{A,B})\le v_m^{A,B}(\vartheta_\star^{A,B},\lambda_\star^{A,B})$. So, we conclude that $(\vartheta_m^{A,B},\lambda_m^{A,B})\in\mathcal{V}_\varepsilon$ for $m\ge m_\alpha$. Therefore, $(\vartheta_m^{A,B},\lambda_m^{A,B})$ converges a.s. to $(\vartheta_\star^{A,B},\lambda_\star^{A,B})$. ■

The above proof essentially relies on the strong convexity of the function $v^{A,B}$ to prove that the pair $(\vartheta_m^{A,B},\lambda_m^{A,B})$ converges to the solution of the original problem. This proof can be adapted when the objective function is only convex and not strictly convex but tends to infinity at infinity.

3.3 The adaptive Monte Carlo estimator

In this section, we assume to have at hand a sequence of optimal solutions $(\vartheta_m^{A,B}, \lambda_m^{A,B})$ and want to devise an adaptive Monte Carlo method taking advantage of of these parameters through the use of Equation (3.7). We propose the following two stages algorithm.

Algorithm 3.3.1

First stage *Generate a sequence $(G^j)_{j=1,\dots,m}$ of i.i.d random vector with the standard normal distribution in $\mathbb{R}^d$ and a sequence $(N^j = (N_1^j, \dots, N_p^j))_{j=1,\dots,m}$ of i.i.d Poisson random vectors with parameter μ. Define*

$$v_m^{A,B}(\vartheta, \lambda) = \frac{1}{m} \sum_{j=1}^{m} f(G^j, N^j)^2 \, \mathrm{e}^{-A\vartheta \cdot G^j + \frac{|A\vartheta|^2}{2}} \prod_{i=1}^{p} \mathrm{e}^{(B\lambda)_i - \mu_i} \left(\frac{\mu_i}{(B\lambda)_i} \right)^{N_i^j}. \tag{3.19}$$

Compute

$$(\vartheta_m, \lambda_m) = \operatorname*{argmin}_{(\vartheta,\lambda) \in \mathbb{R}^{d'} \times \mathbb{R}_+^{*\,p'}} v_m^{A,B}(\vartheta, \lambda).$$

Second stage: *Generate a sequence $(\bar{G}^j)_{j=1,\dots,n}$ of i.i.d random vector with the standard normal distribution in $\mathbb{R}^d$ and a sequence $(\bar{N}^j)_{j=1,\dots,n}$ of i.i.d Poisson random vectors with parameter $B\lambda_m$. Define*

$$M_{n,m}^{A,B} = \frac{1}{n} \sum_{j=1}^{n} f(\bar{G}^j + A\vartheta_m, \bar{N}^j) \, \mathrm{e}^{-A\vartheta_m \cdot \bar{G}^j - \frac{|A\vartheta_m|^2}{2}} \prod_{i=1}^{p} \mathrm{e}^{(B\lambda_m)_i - \mu_i} \left(\frac{\mu_i}{(B\lambda_m)_i} \right)^{\bar{N}_i^j}. \tag{3.20}$$

Note that nothing is said on the dependency structure between $(G^j, N^j)_{1 \le j \le m}$ and $(\bar{G}^j, \bar{N}^j)_{1 \le j \le n}$. In the following, we will study two different cases. First, we assume that conditionally on λ_m, $(G^j, N^j)_{1 \le j \le m}$ and $(\bar{G}^j, \bar{N}^j)_{1 \le j \le n}$ are independent. Second, we prove that in the purely Gaussian case, we can take $G^j = \bar{G}^j$ for all j.

3.3.1 SLLN and CLT in the independent case

In this section, we assume that conditionally on λ_m, $(G^j, N^j)_{1 \le j \le m}$ and $(\bar{G}^j, \bar{N}^j)_{1 \le j \le n}$ are independent. The conditional independence between the two stages combined with Lemma 3.2.1 immediately shows that for any fixed m and n, the estimator $M_{n,m}^{A,B}$ is unbiased, ie. $\mathbb{E}[M_{n,m}^{A,B}] = \mathcal{E}$. Conditionally on $(G_j, N_j)_{j=1,\dots,m}$, the terms involved in the sum of Equation (3.20) are i.i.d., hence the standard strong law of large numbers yields that $\lim_{n \to +\infty} M_{n,m}^{A,B} = \mathbb{E}[f(G, N^\mu)]$ a.s. for any fixed m by applying Lemma 3.2.1. Similarly, the central limit theorem applies and we can state the following result.

Proposition 3.3.2 *For any fixed m, $M_{n,m}^{A,B}$ converges a.s. to $\mathcal{E}$ as n goes to infinity and moreover conditionally on (ϑ_m, λ_m), $\sqrt{n}(M_{n,m}^{A,B} - \mathcal{E}) \xrightarrow[n \to +\infty]{law} \mathcal{N}(0, v^{A,B}(\vartheta_m, \lambda_m))$.*

This result is not fully satisfactory as from a practical point of view, we would like to let both m and n go to infinity. It is convenient to rewrite $M_{n,m(n)}^{A,B}$ using an auxiliary sequence of i.i.d. random

variables $(\bar{U}_i^j)_{1\leq i\leq p, j\geq 1}$ following the uniform distribution on $[0,1]$ and independent of all the other random variables used so far. If we introduce

$$\tilde{N}_i^j(\lambda) = \sum_{k=0}^{\infty} k \mathbf{1}_{P(\lambda_i;k)\leq U_i^j < P(\lambda_i;k+1)} \quad \text{for all } 1\leq i\leq p,\ 1\leq j$$

where $P(\lambda,\cdot)$ is the cumulative distribution function of the Poisson distribution with parameter λ, then $(\bar{N}^j)_{j=1,\ldots,n} \overset{Law}{=} (\tilde{N}^j(\lambda_{m(n)}))_{j=1,\ldots,n}$. Since for all $k\in\mathbb{N}$, the function $\lambda\in\mathbb{R}^*\longmapsto P(\lambda,k)$ is continuous and decreasing, we get that $\lim_{n\to\infty}\tilde{N}^j(\lambda_{m(n)}) = N^j(\lambda_\star)$ a.s. and for all $\lambda\leq\lambda'$, $\tilde{N}^j(\lambda')<\tilde{N}^j(\lambda)$ where the ordering has to be understood component wise.

We define

$$\tilde{M}_n(\theta,\lambda) = \frac{1}{n}\sum_{j=1}^{n} f(\bar{G}^j+\theta,\tilde{N}^j(\lambda))\,\mathrm{e}^{-\theta\cdot\bar{G}^j-\frac{|\theta|^2}{2}}\prod_{i=1}^{p}\mathrm{e}^{\lambda_i-\mu_i}\left(\frac{\mu_i}{\lambda_i}\right)^{\tilde{N}_i^j(\lambda)}.$$

It is obvious that $M_{n,m(n)}^{A,B} \overset{Law}{=} \tilde{M}_n(A\vartheta_{m(n)}, B\lambda_{m(n)})$.

Theorem 3.3.3 *Let $m:\mathbb{N}\to\mathbb{N}$ be an increasing function tending to infinity. Then, under Assumptions ($\mathcal{H}$3.1), $M_{n,m(n)}^{A,B}$ converges a.s. to $\mathcal{E}$ as n goes to infinity.*

Proof. It is actually sufficient to prove the result for A and B being identity matrices. For the sake of clear notations, when $A=I_d$ and $B=I_p$, we write $M_{n,m(n)}$ instead of $M_{n,m(n)}^{A,B}$.

The proof relies on a localising argument combined with Proposition 7.1.1. For a fixed $\varepsilon>0$, we define $\mathcal{V}_\varepsilon$ by

$$\mathcal{V}_\varepsilon \overset{\Delta}{=} \left\{(\theta,\lambda)\in\mathbb{R}^d\times\mathbb{R}^p\ :\ |(\theta,\lambda)-(\theta_\star,\lambda_\star)|\leq\varepsilon\right\}.$$

We assume that ε is small enough, such that $\mathcal{V}_\varepsilon\subset\mathbb{R}^d\times{\mathbb{R}_+^*}^p$. Thanks the independence of the samples used in the two stages of the algorithm, conditionally on $((G^j,N^j),j\geq 1)$, $M_{n,m}$ writes as a sum of i.i.d random variables and $\mathbb{E}[M_{m,n}]=\mathcal{E}$. Consider the sequence

$$X_{j,m} = \left(f(\bar{G}^j+\theta_m,\bar{N}^j)\,\mathrm{e}^{-\theta_m\cdot\bar{G}^j-\frac{|\theta_m|^2}{2}}\prod_{i=1}^{p}\mathrm{e}^{(\lambda_m)_i-\mu_i}\left(\frac{\mu_i}{(\lambda_m)_i}\right)^{\bar{N}_i^j}-\mathcal{E}\right)\mathbf{1}_{(\theta_m,\lambda_m)\in\mathcal{V}_\varepsilon}.$$

Note that $\frac{1}{n}\sum_{i=1}^n X_{j,m} = (M_{m,n}-\mathcal{E})\mathbf{1}_{(\theta_m,\lambda_m)\in\mathcal{V}_\varepsilon}$. By conditioning w.r.t to (θ_m,λ_m), we easily prove that $\mathbb{E}[X_{j,m}]=0$ and

$$\mathbb{E}\left[(M_{n,m}-\mathcal{E})^2\mathbf{1}_{(\theta_m,\lambda_m)\in\mathcal{V}_\varepsilon}\right]\leq\frac{1}{n}\left(\sup_{(\theta,\lambda)\in\mathcal{V}_\varepsilon}v(\theta,\lambda)-\mathcal{E}^2\right).$$

Applying Proposition 7.1.1 proves that $(M_{n,m(n)}-\mathcal{E})\mathbf{1}_{(\theta_{m(n)},\lambda_{m(n)})\in\mathcal{V}_\varepsilon}$ converges to zero a.s. Since, $(\theta_{m(n)},\lambda_{m(n)})\to(\theta_\star,\lambda_\star)$ a.s., we deduce that $M_{n,m(n)}\to\mathcal{E}$ a.s. when n goes to infinity. ■

Theorem 3.3.4 *Let $m:\mathbb{N}\to\mathbb{N}$ be an increasing function of n tending to infinity. Assume that*

($\mathcal{H}$3.3) *i. for all $k\in\mathbb{N}^p$, the function $g\in\mathbb{R}^d\longmapsto f(g,k)$ is continuous;*

ii. there exists a compact neighbourhood $\mathcal{V}$ of $(\vartheta_\star,\lambda_\star)$ included in $\mathbb{R}^{d'}\times{\mathbb{R}_+^*}^{p'}$ and $\eta>0$ such that $\sup_{(\vartheta,\lambda)\in\mathcal{V}}\mathbb{E}\left[|f(\bar{G}+A\vartheta,\tilde{N}^1(B\lambda))|^{2(1+\eta)}\right]<\infty$.

Then, under Assumptions ($\mathcal{H}$3.1) and ($\mathcal{H}$3.2),

$$\sqrt{n}(\tilde{M}_n(A\vartheta_{m(n)}, B\lambda_{m(n)}) - \mathcal{E}) \xrightarrow[n\to+\infty]{law} \mathcal{N}(0, v^{A,B}(\vartheta_\star, \lambda_\star) - \mathcal{E}^2).$$

In [L-5], we proved the result under more stringent assumptions. Here, we manage to relax the assumption $\mathbb{E}\left[\sup_{(\vartheta,\lambda)\in\mathcal{V}} |f(\bar{G} + A\vartheta, \tilde{N}^1(B\lambda))|^{2(1+\eta)}\right] < \infty$ and replace it by $\sup_{(\vartheta,\lambda)\in\mathcal{V}} \mathbb{E}\left[|f(\bar{G} + A\vartheta, \tilde{N}^1(B\lambda))|^{2(1+\eta)}\right] < \infty$. Putting the supremum outside of the expectation is a tremendous improvement in practical applications. Moreover, we required that $m(n) \sim n^\beta$ with $\beta > 0$, which turns out to be unnecessary. We explain the general methodology of the proof for A and B being identity matrices under these relaxed assumptions.

Proof.

$$\sqrt{n}(\tilde{M}_n(\theta_{m(n)}, \lambda_{m(n)}) - \mathcal{E}) = \sqrt{n}(\tilde{M}_n(\theta_\star, \lambda_\star) - \mathcal{E}) + \sqrt{n}(\tilde{M}_n(\theta_{m(n)}, \lambda_{m(n)}) - \tilde{M}_n(\theta_\star, \lambda_\star))$$

From the standard central limit theorem, $\sqrt{n}(\tilde{M}_n(\theta_\star, \lambda_\star) - \mathcal{E}) \xrightarrow[n\to+\infty]{law} \mathcal{N}(0, v(\theta_\star, \lambda_\star) - \mathcal{E}^2)$. Therefore, it is sufficient to prove that $\sqrt{n}(\tilde{M}_n(\theta_{m(n)}, \lambda_{m(n)}) - \tilde{M}_n(\theta_\star, \lambda_\star)) \xrightarrow[n\to+\infty]{Pr} 0$. Let $\varepsilon > 0$.

$$\begin{aligned}\mathbb{P}\left(\sqrt{n}\left|\tilde{M}_n(\theta_{m(n)}, \lambda_{m(n)}) - \tilde{M}_n(\theta_\star, \lambda_\star)\right| > \varepsilon\right) &\le \mathbb{P}(m(n)^{1/4}\left|(\theta_{m(n)}, \lambda_{m(n)}) - (\theta_\star, \lambda_\star)\right| > 1) \\ &+ \frac{n}{\varepsilon^2}\mathbb{E}\left[\left|\tilde{M}_n(\theta_{m(n)}, \lambda_{m(n)}) - \tilde{M}_n(\theta_\star, \lambda_\star)\right|^2 \mathbf{1}_{\left|(\theta_{m(n)}, \lambda_{m(n)}) - (\theta_\star, \lambda_\star)\right| \le m(n)^{-1/4}}\right]. \quad (3.21)\end{aligned}$$

We deduce from Proposition 3.2.5, that $\mathbb{P}(m(n)^{1/4}\left|(\theta_{m(n)}, \lambda_{m(n)}) - (\theta_\star, \lambda_\star)\right| > 1) \longrightarrow 0$. We introduce

$$Q(\theta, \lambda) = \mathrm{e}^{-\theta\cdot\bar{G}^1 - \frac{|\theta|^2}{2}} \prod_{i=1}^{p} \mathrm{e}^{\lambda_i - \mu_i}\left(\frac{\mu_i}{\lambda_i}\right)^{\tilde{N}_i^1(\lambda)}.$$

Conditionally on $(\theta_{m(n)}, \lambda_{m(n)})$, $\tilde{M}_n(\theta_{m(n)}, \lambda_{m(n)}) - \tilde{M}_n(\theta_\star, \lambda_\star)$ is a sum of i.i.d. centered random variables. Then, it is sufficient to monitor differences as

$$\left|f(\bar{G}^1 + \theta_\star, \tilde{N}^1(\lambda_\star))Q(\theta_\star, \lambda_\star) - f(\bar{G}^1 + \theta_{m(n)}, \tilde{N}^1(\lambda_{m(n)}))Q(\theta_{m(n)}, \lambda_{m(n)})\right|^2$$

on the set $\left\{\left|(\theta_{m(n)}, \lambda_{m(n)}) - (\theta_\star, \lambda_\star)\right| \le m(n)^{-1/4}\right\}$, which is a subset of $\mathcal{V}$ for large enough n. Assumption ($\mathcal{H}$3.3-*i)* enables us to prove that it goes to 0 a.s. when n tends to infinity. Let $\delta < \eta$, the conditional independence combined with Hölder's inequality yields that for large enough n

$$\begin{aligned}&\mathbb{E}\left[\left|f(\bar{G}^1 + \theta_{m(n)}, \tilde{N}^1(\lambda_{m(n)}))Q(\theta_{m(n)}, \lambda_{m(n)})\right|^{2(1+\delta)} \mathbf{1}_{(\theta_{m(n)}, \lambda_{m(n)})\in\mathcal{V}}\right] \\ &\le \sup_{(\theta,\lambda)\in\mathcal{V}} \mathbb{E}\left[\left|f(\bar{G}^1 + \theta, \tilde{N}^1(\lambda))Q(\theta, \lambda)\right|^{2(1+\delta)}\right] \\ &\le \sup_{(\theta,\lambda)\in\mathcal{V}} \mathbb{E}\left[\left|f(\bar{G}^1 + \theta, \tilde{N}^1(\lambda))\right|^{2(1+\eta)}\right]^{\frac{1+\delta}{1+\eta}} \mathbb{E}\left[Q(\theta, \lambda)^{\frac{2(1+\delta)(1+\eta)}{\eta-\delta}}\right]^{\frac{\eta-\delta}{1+\eta}}.\end{aligned}$$

Using ($\mathcal{H}$3.3-*ii)*, we deduce that

$$\sup_n \mathbb{E}\left[\left|f(\bar{G}^1 + \theta_{m(n)}, \tilde{N}^1(\lambda_{m(n)}))Q(\theta_{m(n)}, \lambda_{m(n)})\right|^{2(1+\delta)}\right] < \infty,$$

which proves the uniform integrability of the family $(|f(\bar{G}^1 + \theta_{m(n)}, \tilde{N}^1(\lambda_{m(n)}))Q(\theta_{m(n)}, \lambda_{m(n)})|^2)_n$. Then, we deduce that the second term in (3.21) tends to zero. ■

3.3.2 Recycling the samples in the Gaussian case

In this section, we focus on the purely Gaussian framework, which writes

$$\mathcal{E} = \mathbb{E}\left[f(G+\vartheta)\,\mathrm{e}^{-A\vartheta\cdot G-\frac{|A\vartheta|^2}{2}}\right], \quad \forall\,\vartheta\in\mathbb{R}^{d'},\ \forall\,A\in\mathcal{M}_{d\times d'}.$$

The variance associated to the parametrized representation is given by $v(\theta)-\mathcal{E}^2$ where

$$v(\vartheta) = \mathbb{E}\left[f(G)^2\,\mathrm{e}^{-A\vartheta\cdot G+\frac{|A\vartheta|^2}{2}}\right]. \tag{3.22}$$

In [L-12], we considered Algorithm 3.3.1 with the same samples in both stages instead of sampling conditionally independent random vectors between the two stages. This is easily done and natural since a normal random vector X with mean vector θ can be written as $X=\theta+G$ where G is a standard normal random vector. No such simple relation exists for the Poisson distribution to link a Poisson random variable with parameter μ to one with parameter λ, which explains why we have just independent samples in Section 3.3.1.

Algorithm 3.3.5 *Generate a sequence $(G^j)_{j=1,\dots,m\vee n}$ of i.i.d random vectors following the standard normal distribution in $\mathbb{R}^d$.*

First stage *Define*

$$v_m^A(\vartheta) = \frac{1}{m}\sum_{j=1}^{m} f(G^j)^2\,\mathrm{e}^{-A\vartheta\cdot G^j+\frac{|A\vartheta|^2}{2}}. \tag{3.23}$$

Compute

$$\vartheta_m = \operatorname*{argmin}_{\vartheta\in\mathbb{R}^{d'}} v_m^A(\vartheta).$$

Second stage: *Compute*

$$M_{n,m}^A = \frac{1}{n}\sum_{j=1}^{n} f(G^j+A\vartheta_m)\,\mathrm{e}^{-A\vartheta_m\cdot G^j-\frac{|A\vartheta_m|^2}{2}}. \tag{3.24}$$

Theorem 3.3.6 *Let $m:\mathbb{N}\to\mathbb{N}$ be an increasing function tending to infinity. Assume that $\mathbb{P}(|f(G)|>0)>0$ and that there exists $\gamma>0$ s.t. $\mathbb{E}[|f(G)|^{2+\gamma}]<\infty$. If the function f is continuous and*

$$\forall\,K>0,\ \mathbb{E}\left[\sup_{|\theta|\le K}|f(G+\theta)|\right]<\infty, \tag{3.25}$$

then $M_{n,m(n)}^A$ converges a.s. to $\mathbb{E}[f(G)]$ as n goes to infinity.

Compared to the case in which the samples used in both stages are independent, more stringent assumptions are required to ensure the a.s. convergence. In Theorem 3.3.3, which is the counterpart of Theorem 3.3.6 for the independent case, we did not require any continuity or conditions of the type of (3.25). As before, we introduce

$$\tilde{M}_n(\theta) = \frac{1}{n}\sum_{j=1}^{n} f(\bar{G}^j+\theta)\,\mathrm{e}^{-\theta\cdot\bar{G}^j-\frac{|\theta|^2}{2}}.$$

The proof of Theorem 3.3.6 ensues from the locally uniform strong law of large numbers for $\tilde{M}_n$.

Theorem 3.3.7 *Assume that* $\mathbb{P}(|f(G)|>0)>0$ *and that there exists* $\gamma>0$ *s.t.* $\mathbb{E}[|f(G)|^{4+\gamma}]<\infty$. *Assume that the function* f *admits a decomposition* $f=f_1+f_2$ *with* f_1 *of class* C^1 *satisfying*

$$\forall K>0,\ \mathbb{E}\left[\sup_{|\theta|\leq K}|f_1(\theta+G)|+\sup_{|\theta|\leq K}|\nabla f_1(\theta+G)|\right]<\infty$$

and f_2 *satisfies*

$$\exists\beta\in[0,2),\exists\lambda>0,\ \forall x,y\in\mathbb{R}^d,\ |f_2(x)-f_2(y)|\leq\lambda e^{|x|^\beta\vee|y|^\beta}|x-y|^\alpha$$

for $\alpha\in\left(\frac{\sqrt{d'^2+8d'}-d'}{4},1\right]$. *Then, for any increasing function* $m:\mathbb{N}\to\mathbb{N}$ *s.t.* $m(n)\sim n^\delta$ *for* $\delta>\frac{d'}{\alpha(d'+2\alpha)}$,

$$\sqrt{n}(M^A_{n,m(n)}-\mathcal{E})\ \xrightarrow[n\to\infty]{law}\mathcal{N}\left(0,v^A(\vartheta_\star)-\mathcal{E}^2\right).$$

By the standard central limit theorem, $\sqrt{n}(M_n(A\vartheta_\star)-\mathcal{E})\xrightarrow[n\to\infty]{law}\mathcal{N}\left(0,v^A(\vartheta_\star)-\mathcal{E}^2\right)$. As a consequence, it is enough to check that $\sqrt{n}(M_n(A\vartheta_{m(n)})-M_n(A\vartheta_\star))\xrightarrow{Pr}0$. For a precise statement of the different results and their proofs, we refer to [L-12].

Remark 3.3.8 (Comparison with the results of Section 3.3.1) In this section, we proved that in the Gaussian case the same samples could be used to first compute an approximation of the optimal importance sampling parameter and then to compute the Monte Carlo estimator. This leads to an estimator with no remarkable structure, which nonetheless satisfies a central limit theorem with optimal limiting variance. When using the same samples, the convergence results require more stringent assumptions. Instead of moment assumptions on the function f for the independent case, we need conditions as $\forall\ K>0,\ \mathbb{E}\left[\sup_{|\theta|\leq K}f(G+\theta)\right]<\infty$ to prove the convergence and C^1 regularity along with $\forall K>0,\ \mathbb{E}\left[\sup_{|\theta|\leq K}|\nabla f_1(\theta+G)|\right]<\infty$ to obtain the asymptotic normality.

3.3.3 Practical implementation

The difficult part of Algorithm 3.3.1 is the numerical computation of the minimizing pair (θ_m,λ_m). The efficiency of the optimization algorithm depends very much on the magnitude of the smallest eigenvalue of $\nabla^2 v$. For the sake of clearness, we present the methodology only in the pure Gaussian case, the mixed Gaussian Poisson case is detailed in [L-5]. From (3.11), we deduce that the smallest eigenvalue of $\nabla^2 v$ is larger than

$$\mathbb{E}\left[f(G)^2\,\mathrm{e}^{-\theta\cdot G+\frac{|\theta|^2}{2}}\right].$$

This lower bound depends on the function f whereas we would rather have a uniform lower bound. We advice to rewrite ∇v as

$$\nabla v(\theta,\lambda)=\mathbb{E}\left[\theta f(G)^2\,\mathrm{e}^{-\theta\cdot G+\frac{|\theta|^2}{2}}\right]-\mathbb{E}\left[Gf(G)^2\,\mathrm{e}^{-\theta\cdot G+\frac{|\theta|^2}{2}}\right]$$

Hence, $\theta^\star$ can be seen as the root of

$$\nabla u(\theta)=\theta-\frac{\mathbb{E}\left[Gf(G)^2\,\mathrm{e}^{-\theta\cdot G}\right]}{\mathbb{E}\left[f(G)^2\,\mathrm{e}^{-\theta\cdot G}\right]}$$

with $u(\theta) = \frac{|\theta|^2}{2} + \log \mathbb{E}\left[f(G)^2 \, \mathrm{e}^{-\theta \cdot G}\right]$. The Hessian matrix of u is given by

$$\nabla^2 u(\theta) = I_d + \text{"a positive semi–definite matrix"} \geq I_d.$$

Our numerical experiments advocate the use of u instead of v to speed up the computation of $\theta_\star$.

Using this new expression, we implement Algorithm 3.3.10 to construct an approximation x_m^k of (θ_m, λ_m). Since u_m is strongly convex, for any fixed m, x_m^k converges to (θ_m, λ_m) when k goes to infinity. The descent direction d_m^k at step k should be computed as the solution of a linear system. There is no point in computing the inverse of $\nabla^2 u_m(x_m^k)$, which would be computationally much more expensive.

Remark 3.3.9 (Remarks on the implementation) From a practical point of view, ε should be chosen reasonably small $\varepsilon \approx 10^{-6}$. This algorithm converges very quickly and, in most cases, less than 5 iterations are enough to get a very accurate estimate of (θ_m, λ_m), actually within the $\varepsilon-$error. Since the points at which the function f is evaluated remain constant through the iterations of Newton's algorithm, the values $(f^2(G^j, N^j))_{1 \leq j \leq m}$ should be precomputed before starting the optimization algorithm, which considerably speeds up the whole process. The Hessian matrix of our problem is easily tractable so there is no point in using quasi Newton methods.

Choose an initial value $x_m^0 \in \mathbb{R}^{d+p}$.
$k \leftarrow 1$
while $\left|\nabla u_m(x_m^k)\right| > \varepsilon$ **do**
  Compute d_m^k such that $(\nabla^2 u_m(x_m^k)) d_m^k = -\nabla u_m(x_m^k)$
  $x_m^{k+1/2} = x_m^k + d_m^k$
  for $i = 1 : d + p$ **do**
    if $x_m^{k+1/2}(i) > 0$ **then**
      $x_m^{k+1}(i) = x_m^{k+1/2}(i)$
    else
      $x_m^{k+1}(i) = \frac{x_m^k(i)}{2}$
    end
  end
  $k \leftarrow k + 1$
end

Algorithm 3.3.10: Projected Newton's algorithm

3.4 Application to option pricing

We will apply our methodology to two different classes of jump processes: jump diffusion processes and stochastic volatility processes with jumps, in this latter case the volatility itself may also jump.

We consider a filtered probability space $(\Omega, (\mathcal{F}_t)_{0 \leq t \leq T}, \mathbb{P})$ with a finite time horizon $T > 0$ and N financial assets. We define on this space a Brownian motion W with values in $\mathbb{R}^N$ and $N + 1$ independent Poisson processes $(N^1, \ldots, N^{N+1})$ with constant intensities $\mu^1, \ldots, \mu^{N+1}$. We also consider $(N + 1)$ independent sequences $(Y_j^i)_{j \geq 1}$ for $i = 1 \ldots N + 1$ of i.i.d. real valued random variables with common law denoted Y in the following. The Poisson processes, the Brownian motions

and the sequences $(Y_j^i)_j$ are supposed to be independent of each other. Actually, we are interested in considering the compound Poisson process associated to the Poisson process N^i and to the jump sequences Y^i for $i = 1, \dots, N+1$.

3.4.1 Black–Scholes model with jumps

In this class of models, we assume that the log-prices evolve according to the following equation

$$X_t^i = \left(\beta^i - \frac{(\sigma^i)^2}{2}\right) t + \sigma^i L^i W_t + \sum_{j=1}^{N_t^i} Y_j^i + \sum_{j=1}^{N_t^{N+1}} Y_j^{N+1} \tag{3.26}$$

where $\beta = (\beta^1, \dots, \beta^N)^*$ is the drift vector and $\sigma = (\sigma^i, \dots, \sigma^N)^*$ the volatility vector. The row vectors L_i are such that the matrix $L = (L^1; \dots; L^N)$ verifies that $\Gamma = LL^*$ is a symmetric definite positive matrix with unit diagonal elements. The matrix Γ embeds the covariance structure of the continuous part of the model. We have also chosen to take into account in the model the possibility to have simultaneous jumps which explains the extra jump term $\sum_{j=1}^{N_t^{N+1}} Y_j^{N+1}$ common to all the underlying assets. This common jump term corresponds to the systemic risk of the market.

From (3.26), we deduce that the prices at time t $S_t^i = \mathrm{e}^{X_t^i}$ are defined by

$$S_t^i = S_0^i \exp\left\{\left(\beta^i - \frac{(\sigma^i)^2}{2}\right) t + \sigma^i L^i W_t\right\} \prod_{j=1}^{N_t^i} \mathrm{e}^{Y_j^i} \prod_{j=1}^{N_t^{N+1}} \mathrm{e}^{Y_j^{N+1}}$$

which corresponds for each asset to a one dimensional Merton model with intensity $\mu^i + \mu^{N+1}$ when the Y_j^i are normally distributed.

As we assumed that $\mathbb{P}$ was the martingale measure associated to the risk free rate $r > 0$ supposed to be deterministic, the processes $(\mathrm{e}^{-rt} S_t)_t$ must be martingales under $\mathbb{P}$. This martingale condition imposes that for every $i = 1, \dots, N$,

$$\beta^i = r - (\mu^i \mathbb{E}[Y^i] + \mu^{N+1} \mathbb{E}[Y^{N+1}]).$$

In the following, β_i will always stand for this quantity.

Remark 3.4.1 In the one dimensional case, ie. when $N = 1$, we only consider a single compound Poisson process as the systemic risk jump term becomes irrelevant. Hence, the log-price in dimension one follows

$$X_t = \left(\beta - \frac{\sigma^2}{2}\right) t + \sigma W_t + \sum_{j=1}^{N_t} Y_j.$$

For the sake of clearness, we will not treat the one dimensional case separately in the following, even though the practical one dimensional implementation relies on a single Poisson process. So, we will always consider that the Poisson process has values in $\mathbb{R}^{N+1}$.

In the numerical examples, we will need to discretize the multi dimensional price process on a time grid $0 = t_0 < t_1 < \cdots < t_J = T$. We will assume that this time grid is regular and given by $t_j = \frac{jT}{J}$, $j = 0, \dots, J$.

The Merton jump diffusion model. The Merton model corresponds to the particular choice of a normal distribution for the variables (Y^i), $Y^i \sim \mathcal{N}(\alpha, \delta)$ where $\alpha \in \mathbb{R}$ and $\delta > 0$. In this framework, the jump sizes in the price follow a log normal distribution.

The Kou model. In the Kou model [64], the variables Y^i follow an asymmetric exponential distribution with density

$$p^i \mu_+^i \, \mathrm{e}^{-\mu_+^i x} \, \mathbf{1}_{x>0} + (1-p)^i \mu_-^i \, \mathrm{e}^{\mu_-^i x} \, \mathbf{1}_{x<0}$$

where $p^i \in [0,1]$ is the probability of a positive jump for the $i-th$ component and the variables $\mu_+^i > 0, \mu_-^i > 0$ govern the decay of each exponential part.

3.4.2 Stochastic volatility models with jumps

In this section, we consider the stochastic volatility model developed by [7, 8], in which the volatility process is a non Gaussian Ornstein Uhlenbeck process driven by a compound Poisson process.

We consider that the log-prices satisfy for $i = 1, \ldots, N$

$$dX_t^i = (a^i - \sigma^i/2)dt + \sqrt{\sigma_{t^-}^i} dW_t^i + \psi^i dZ^i_{\kappa^i t} + \psi^{N+1} dZ^{N+1}_{\kappa^{N+1} t}$$

where $a \in \mathbb{R}^N$, $\psi \in \mathbb{R}^{N+1}$ has non-positive components which account for the positive leverage effect, Z is a $(N+1)$-dimensional Lévy process defined by $Z_t^i = \sum_{k=1}^{N_t^i} Y_k^i$ for $i = 1, \ldots, N+1$ and the squared volatility process $(\sigma_t)_t$ is a Lévy driven Ornstein Uhlenbeck processus

$$N\sigma_t^i = -(\kappa^i + \kappa^{N+1})\sigma_t^i dt + dZ^i_{\kappa^i t} + dZ^{N+1}_{\kappa^{N+1} t}.$$

For the squared volatility process to remain positive, we assume that the components of Z only jumps upward, which means that the random variables Y_j^i are non-negative.

More specifically, the jump sequence Y^i is i.i.d with the exponential distribution with parameter $\beta^i > 0$ for $i = 1, \ldots, N+1$. The drift vector a is chosen such that the discounted prices are martingales under $\mathbb{P}$. A straight computation shows that we need to set

$$a^i = r - \psi^i \frac{\kappa^i \mu^i}{\beta^i - \psi^i} - \psi^{N+1} \frac{\kappa^{N+1} \mu^{N+1}}{\beta^{N+1} - \psi^{N+1}}, \quad \text{for } i = 1, \ldots, N$$

to ensure the martingale property of $(\mathrm{e}^{-rt} \exp X_t)_t$.

The extra Poisson process giving raise to the term dZ^{N+1} in the dynamics of X and σ accounts for modeling a systemic risk. When Z^{N+1} jumps, all the volatilities and possibly all the assets (when there is a leverage effect) jump together. This parametrization of multi-dimensional stochastic volatility models with jumps corresponds to Section 5.3 of [9].

In the following, we compare the efficiencies of several different approaches based on the theoretical part of the paper in the context of option pricing with jumps. The problem always boils down to computing the expectation of a function of a jump diffusion process.

3.4.3 Several importance sampling approaches

When dealing with jump diffusions, importance sampling can act only the Brownian part — referred to hereafter as *Gaussian importance sampling* with an optimal variance denoted *VarG*, or only on the Poisson part — referred to as *Poisson importance sampling* with an optimal variance denoted *VarP*,

or on both of them at the same time. This last approach is named *Gaussian+Poisson importance sampling* and leads to an optimal variance denoted *VarGP*. When only playing with the Gaussian part, one can either use the same samples to compute the optimal $\theta_\star$ and the Monte Carlo estimator as explained in Section 3.3.2 or use independent set of samples. Both approaches have the same computational cost and leads to the same variance reduction.

For each of the three methods, we consider two approaches.

Full importance sampling. The first approach consists in allowing to optimize the parameters per time steps, this means that $d = d' = N \times J$ and $p = p' = (N+1) \times J$. In this setting, the matrices A and B are identity matrices. This is the most general approach, but the dimension of the optimization problem linked to the variance minimization increases linearly in the number of time steps J and in the number of assets d. Then, it is worth trying to find a vector subspace with smaller dimension in which the optimal variance is close the global minimum.

Reduced importance sampling. The idea of reducing the dimension of the problem is to search for the parameter (θ, λ) in the subspace $\{(A\vartheta, B\lambda) \ : \ \vartheta \in \mathbb{R}^{d'}, \ \lambda \in {\mathbb{R}_+^*}^{p'}\}$ where $A \in \mathbb{R}^{d \times d'}$ is a matrix with rank $d' \leq d$ and $B \in {\mathbb{R}_+^*}^{p \times p'}$ a matrix with rank $p' \leq p$.

We choose $d' = N$, $p' = N + 1$ and

$$A_{(j-1)N+i,i} = \sqrt{t_j - t_{j-1}}, \quad B_{(j-1)(N+1)+k,k} = t_j - t_{j-1}$$

for $j = 1, \ldots, J$, $i = 1, \ldots, N$ and $k = 1, \ldots, N+1$, all the other coefficients of A and B being zero. This choice corresponds to adding a linear drift to the Brownian motion and to keeping the Poisson intensity time independent.

3.4.4 Numerical experiments

We compare the different importance sampling approaches on four different financial derivatives: the first two examples are path-dependent single asset options while the last two examples are basket option with or without barrier monitoring. To compare the different strategies, we have decided to fix the number of samples for the Monte Carlo part, which implies that their accuracies only depend on their variances, which we will compare in different examples. To determine which method is best, it is convenient to compute their efficiencies defined as the ratio of the variance divided by the CPU time.

In all the following examples, we use the same number of samples for the approximation of the optimal importance sampling parameters and for the Monte Carlo computation, ie. $m(n) = n$.

Asian option. We consider a discretely monitored Asian option with payoff

$$\left(\frac{1}{J} \sum_{i=1}^{J} S_{t_i} - K \right)_+ .$$

Our tests on one dimensional Asian options (see Tables 3.1 and 3.2) show that the *Poisson* and *Gaussian+Poisson* importance sampling methods perform generally better than the pure *Gaussian* importance sampling approach but they also require a longer computational time. When taking into account this extra computational times along with the variance reduction we notice that the *Poisson*

and *Gaussian+Poisson* importance sampling methods yield the same efficiency for the Merton model (see Table 3.1). For the BNS model (see Table 3.2), the mixed *Gaussian+Poisson* importance sampling approach achieves a better variance reduction than the two other methods for a comparable computational time. By closely looking at the CPU times of the different strategies, it clearly appears that the reduced approach shows the better efficiency and should be used in practice.

	Strike	Price	Var	VarG	VarP	VarGP
Full	90	17.88	2639	2395	636	529
Reduced		17.88	2639	2640	839	752
Full	100	14.37	2750	2624	720	622
Reduced		14.37	2750	2624	552	470
Full	110	12.11	2327	2301	470	420
Reduced		12.11	2327	2301	676	585

Table 3.1: Discrete Asian option in dimension 1 in the Merton model with $S_0 = 100$, $r = 0.05$, $\sigma = 0.25$, $\mu = 1$, $\alpha = 0.5$, $\delta = 0.2$, $T = 1$, $J = 12$ and $n = 50000$. The CPU time for the crude Monte Carlo approach is 0.08. The CPU times for the full importance sampling approach are $(0.21, 0.28, 0.39)$ and for the reduced approach $(0.20, 0.21, 0.26)$.

	Strike	Price	Var	VarG	VarP	VarGP
Full	90	11.85	63	22.7	50	13.3
Reduced		11.85	63	28.7	52.7	22.1
Full	100	3.96	47	19	29.7	9.4
Reduced		3.96	47	22	33	14.7
Full	110	0.92	19	7.8	9	3.5
Reduced		0.92	19	10	11.1	5.56

Table 3.2: Discrete Asian option in dimension 1 in the BNS model with $S_0 = 100$, $r = 0.05$, $\lambda_0 = 0.01$, $\mu = 1$, $\kappa = 0.5474$, $\beta = 18.6$, $T = 1$, $J = 12$ and $n = 50000$. The CPU time for the crude Monte Carlo approach is 0.13. The CPU times for the full importance sampling approach are $(0.36, 0.52, 0.93)$ and for the reduced approach $(0.29, 0.29, 0.33)$.

Barrier option. We consider a discrete monitoring barrier option with payoff

$$(S_T - K)_+ \times \mathbf{1}_{\forall 1 \leq j \leq J,\, S_{t_j} < U}$$

where U is the upper barrier.

Basket option. We consider a basket option on 10 assets with payoff

$$\left(\sum_{i=1}^{N} \omega^i S_T^i - K \right)_+$$

where the vector $\omega \in \mathbb{R}^N$ describes the weight of each asset in the basket.

The experiments on the one dimensional barrier option (see Table 3.3) lead to very similar conclusions regarding the efficiencies of the different approaches. Roughly speaking, the *Gaussian* approach does not bring any variance reduction but costs 2.5 times the CPU times of the crude Monte Carlo

	Strike	Price	Var	VarG	VarP	VarGP
Full	90	17.88	2639	2395	636	529
Reduced		17.88	2639	2640	839	752
Full	100	14.37	2750	2624	720	622
Reduced		14.37	2750	2624	552	470
Full	110	12.11	2327	2301	470	420
Reduced		12.11	2327	2301	676	585

Table 3.3: Discrete barrier option in dimension 1 in the Merton model with $S_0 = 100$, $r = 0.05$, $\sigma = 0.2$, $\mu = 0.1$, $\alpha = 0$, $\delta = 0.1$, $T = 1$, $J = 12$, $U = 140$ and $n = 50000$. The CPU time for the crude Monte Carlo approach is 0.08. The CPU times for the full importance sampling approach are $(0.22, 0.26, 0.37)$ and for the reduced approach $(0.20, 0.20, 0.23)$.

Strike	Price	Var	VarG	VarP	VarGP
-10	10.61	112	85	66	48
0	3.66	85	66	33	25
10	1.17	111	52	12	10

Table 3.4: Basket option in dimension $N = 10$ in the Merton model with $S_0^i = 100$, $r = 0.05$, $\sigma^i = 0.2$, $\mu^i = 0.1$, $\alpha^i = 0.3$, $\delta^i = 0.2$, $\rho = 0.3$, $T = 1$, $\omega^i = \frac{1}{N}$ for $i = 1, \ldots, N/2$, $\omega^i = -\frac{1}{N}$ for $i = N/2+1, \ldots, N$ and $n = 50000$. The CPU time for the crude Monte Carlo approach is 0.06. The CPU times for the importance sampling approach are $(0.17, 0.20, 0.32)$.

approach. The *Poisson* and *Gaussian+Poisson* importance sampling approaches do provide impressive variance reductions for equivalent computational times at least in the reduced size approach. The improvement of the optimal variance obtained by the full size approaches does not look sufficient to counter balance the extra computational time. Actually, the reduced size approaches show far better efficiencies.

Strike	Price	Var	VarG	VarP	VarGP
-10	10.21	60	41	48	29
0	3.35	30	21	22	13
10	0.68	8.3	5.9	5.2	2.8

Table 3.5: Basket option in dimension $N = 10$ in the Merton model with $S_0^i = 100$, $r = 0.05$, $\sigma^i = 0.2$, $\mu^i = 1$, $\alpha^i = 0.1$, $\delta^i = 0.01$, $\rho = 0.3$, $T = 1$, $\omega^i = \frac{1}{N}$ for $i = 1, \ldots, N/2$, $\omega^i = -\frac{1}{N}$ for $i = N/2+1, \ldots, N$ and $n = 50000$. The CPU time for the crude Monte Carlo approach is 0.06. The CPU times for the importance sampling approach are $(0.17, 0.20, 0.32)$.

Since basket options are not path dependent derivatives, the full and reduced size approaches coincide and we do not distinguish between the two in Tables 3.4 and 3.5. In these tables, we can see that the *Gaussian+Poisson* approach provides better variance reductions that the pure *Poisson* approach, which in turn outperforms the pure *Gaussian* strategy. However, except for out of the money options, the gain brought by the different importance sampling approaches do not compensate the extra computational time in order to keep up with the crude Monte Carlo strategy. This lack of efficiency mainly comes from the very simple form of the payoff, which makes the crude Monte Carlo method very fast.

Multidimensional barrier option. We consider a discrete monitoring down and out barrier option on a basket of assets with payoff

$$\left(\sum_{i=1}^{N} \omega^i S_T^i - K\right)_+ \mathbf{1}_{\forall 1\leq i\leq N,\, \forall 1\leq j\leq J,\, S^i_{t_j} > b^i}$$

where the vector $b \in \mathbb{R}^N$ denotes the lower barrier.

	Strike	Price	Var	VarG	VarP	VarGP
Full	0	0.59	7.00	4.03	4.36	1.98
Reduced		0.59	7.00	3.51	4.36	2.05
Full	-5	1.06	13.33	8.43	9.64	4.79
Reduced		1.06	13.33	8.56	9.81	5.42
Full	-10	1.64	24.26	16.57	18.39	10.57
Reduced		1.64	24.26	16.77	18.96	11.68

Table 3.6: Barrier option in dimension $N = 10$ in the Merton model with $S_0^i = 100$, $r = 0.05$, $\sigma^i = 0.2$, $\mu^i = 1$, $\alpha^i = 0.1$, $\delta^i = 0.01$, $b^i = 80$, $\rho = 0.3$, $T = 1$, $\omega^i = \frac{1}{N}$ for $i = 1, \ldots, N/2$, $\omega^i = -\frac{1}{N}$ for $i = N/2+1, \ldots, N$ and $J = 12$, $n = 50000$. The CPU time for the crude Monte Carlo approach is 0.76. The CPU times for the reduced importance sampling approach are $(1.42, 1.44, 1.52)$ and for the full importance sampling approach they are $(1.53, 2.41, 3.05)$.

	Strike	Price	Var	VarG	VarP	VarGP
Full	100	2.97	36	37.8	16	16
Reduced		2.97	36	36.2	16	16
Full	90	12.52	36	36.4	14.5	14.5
Reduced		12.52	36	36	14.3	14.3
Full	110	1.64	12.1	13.3	6.1	5.5
Reduced		0.80	12.1	12	5.3	5.4

Table 3.7: Barrier option in dimension $N = 5$ in the BNS model with $S_0^i = 100$, $r = 0.05$, $\lambda^i = 0.01$, $\mu^i = 1$, $\kappa^i = 0.54$, $\beta^i = 18.6$, $b^i = 70$, $\rho = 0.2$, $T = 1$, $\omega^i = \frac{1}{N}$ for $i = 1, \ldots, N$ and $J = 12$, $n = 50000$. The CPU time for the crude Monte Carlo approach is 0.52. The CPU times for the reduced importance sampling approach are $(1.06, 1.1, 1.17)$ and for the full importance sampling approach they are $(2.1, 3.8, 10.5)$.

Our last two examples deal with multi-dimensional barrier options with discrete monitoring. The first striking result to notice when looking at Tables 3.6 and 3.7 concerns the huge CPU times of the full approaches which nonetheless do not significantly reduce the variance compared to the reduced size methods. This remark definitely advocates the use of reduced size approaches. The variance is always divided by a factor between 2 and 3, whereas the CPU time is only twice the one of the crude Monte Carlo approach. In the Merton model case (Table 3.6), the *Gaussian+Poisson* approach always provides the best variance reduction for a computational time very close to the other two methods, while the *Poisson* and *Gaussian+Poisson* methods perform similarly in the BNS model (Table 3.7). The efficiency of the pure *Poisson* approach comes from the particular form of the BNS model which includes jumps in the volatility process. These jumps seem to have a larger impact on the overall variance that the Brownian motion itself.

3.5 Conclusion

In this chapter, we have presented adaptive Monte Carlo methods based on importance sampling. Unlike most works on the topic which use *stochastic approximation* to compute the optimal importance sampling parameter, we rely on *sample average approximation*, which basically consists in replacing expectations by sample averages and then perform deterministic optimization on them. We proved that the solution of the approximated problem converges to the solution of the original problem and satisfies a central limit theorem. The computation of the best importance sampling distribution leads to a strongly convex and infinitely differentiable optimization problem. The sample average approximation enables us to make the utmost of these regularity properties by relying on Newton's algorithm with optimal step size, which reveals so efficient while being easy to implement.

The numerical examples presented in Section 3.4 all involved the discretization of a stochastic differential equation. It is well that, in this context, the total error can be split into two terms: the bias term coming from the time discretization of the SDE and the variance term coming from the Monte Carlo method. In this chapter, we focused only on the variance term, but the bias term should also be taken into account. In the next chapter, we present how to couple importance sampling, which only acts on the variance term, with multilevel Monte Carlo methods, which are known for reducing both the bias and the variance.

Chapter 4

Coupling Multilevel Monte Carlo with importance sampling

In this chapter, we present how to couple multilevel Monte Carlo methods with importance sampling, see [L-1].

4.1 Introduction

Expectations involving a stochastic process are often computed using a Monte Carlo method combined with a discretization scheme. For instance, computing an hedging portfolio in finance uses these tools. Generally, the asset price follows a diffusion process $(X_t)_{0\leq t\leq T} \in \mathbb{R}^d$ with a non explicit solution, whose simulation requires a discretization scheme $(X_t^n)_{0\leq t\leq T}$ with $n \in \mathbb{N}^*$ time steps. The error induced by such schemes is called the discretization error or the bias. Then, the valuation of a financial derivative using a Monte Carlo method involves the simulation of N independent samples of X_T^n. These methods are known to converge slowly. In particular, for a given discretization error of order $1/n^\alpha$, for $\alpha > 0$, the optimal choice for the number of samples is given by $N = n^{2\alpha}$. This leads to an overall complexity for the Monte Carlo method of order $n^{3\alpha}$. Nevertheless, a lot of techniques have been developed in the recent years to speed up the method. Kebaier [59] proposed the Statistical Romberg method to generate discretization schemes on two different time grids, using a coarser grid to simulate a crude approximation and a finer one to tune the bias. More recently, Giles [41] generalized the statistical Romberg method and proposed the multilevel Monte Carlo algorithm in a similar approach to Heinrich's multilevel method for parametric integration, see Heinrich [50]. It turns out that for the Euler scheme with a given discretization error of order $1/n^\alpha$, $\alpha > 0$, and for a Lipschitz continuous payoff function, the optimal complexity of the Statistical Romberg and the multilevel Monte Carlo methods are respectively of order $n^{2\alpha+1/2}$ and $n^{2\alpha}(\log n)^2$, which are clearly better than a crude Monte Carlo method. We refer the reader to the extensive literature linked to Multilevel Monte Carlo for more details on the rate of decrease of the mean squared error, see Dereich [35], Giles [41], Giles and Szpruch [42], Giles et al. [43], Heinrich [49], Heinrich and Sindambiwe [51].

The use of multilevel techniques clearly reduces the bias, but in many situations the high variance also brings in a significant inaccuracy, which naturally leads to trying to couple multilevel Monte Carlo with variance reduction techniques. In this work, we focus on importance sampling following the ideas and the methodology developed in Chapter 3. Consider a parametric family $(X_t(\theta))_{0\leq t\leq T}$ driven by a Brownian motion with linear drift $(\theta t)_{0\leq t\leq T}$. Coupling importance sampling and multilevel Monte Carlo can be achieved by minimizing the asymptotic variance of the multilevel method,

which involves a bespoke tangent process defined on an augmented probability space and the gradient of the function whose expectation is to be computed, see Ben Alaya and Kebaier [12]. Hajji [45] used a stochastic algorithm as presented in Chapter 2 to minimize the limiting variance. From our point of view, this approach suffers from a major drawback. It requires some extra implementation to compute the gradient and the auxiliary modified gradient process, which must also be discretized.

To circumvent this difficulty, we use one importance sampling parameter per level and choose it as the minimizer of the variance of the level. This approach preserves the independence structure of the levels, which makes our approach suitable for running on parallel architectures. From a practical point of view, the use of as many importance sampling parameters as the number of levels represents a huge improvement. Our approach is closely related to minimizing the multilevel estimator of the limiting variance.

4.2 The importance sampling framework

Let $(X_t)_{0\leq t\leq T}$ be the solution of

$$dX_t = b(X_t)dt + \sigma(X_t)dW_t, \quad X_0 = x \in \mathbb{R}^d \tag{4.1}$$

where W is a q-dimensional Brownian motion on some given probability space $(\Omega, (\mathcal{F}_t)_{0\leq t\leq T}, \mathbb{P})$ with finite time horizon $T > 0$. The functions $b : \mathbb{R}^d \longrightarrow \mathbb{R}^d$ and $\sigma : \mathbb{R}^d \longrightarrow \mathcal{M}_{d\times d}$ are assumed to be Lipschitz, which ensures the strong existence and uniqueness of a solution to (4.1). In many applications, in particular when pricing financial securities, we are interested in the effective computation by Monte Carlo methods of the quantity $\mathbb{E}[\psi(X_T)]$ for a given function ψ. From a practical point of view, we have to discretize the process X. Let us consider the continuous time Euler approximation X^n with time step $\delta = T/n$ given by

$$dX_t^n = b(X^n_{\eta_n(t)})dt + \sigma(X^n_{\eta_n(t)})dW_t, \quad \eta_n(t) = \lfloor t/\delta \rfloor \delta.$$

It is well known that, under the Lipschitz condition, X^n converges to X in L^p with rate $n^{-1/2}$ (see e.g. Bouleau and Lépingle [19]). The weak error was first studied by Talay and Tubaro [87] who proved that if ψ, b and $(\sigma_j)_{1\leq j\leq q}$ are four times differentiable and together with their derivatives have at most polynomial growth, then $\mathbb{E}[\psi(X_T^n)] - \mathbb{E}[\psi(X_T)] = O(1/n)$. The same result was later extended by Bally and Talay [6] for a measurable function ψ but with a non degeneracy condition of Hörmander's type on the diffusion. In the context of possibly degenerate diffusions, Kebaier [59] showed the rate of convergence can be $1/n^\gamma$, for any $\gamma \in [1/2, 1]$. So, it is worth introducing the following assumption

$$\exists\, \gamma \in [1/2, 1], \exists\, C_\psi(T, \gamma) \in \mathbb{R}, \quad n^\gamma(\mathbb{E}\psi(X_T^n) - \mathbb{E}\psi(X_T)) \to C_\psi(T, \gamma). \tag{4.2}$$

We define the family $(\mathbb{P}_\theta)_{\theta\in\mathbb{R}^q}$ of equivalent probability measures such that for all $t > 0$

$$L_t^\theta = \frac{d\mathbb{P}_\theta}{d\mathbb{P}}_{|\mathcal{F}_t} = \exp\left(\theta \cdot W_t - \frac{1}{2}|\theta|^2 t\right).$$

From Girsanov's theorem, the process $(B_t^\theta = W_t - \theta t)_{t\leq T}$ is a Brownian motion under $\mathbb{P}_\theta$ and moreover if the process $X(\theta)$ is the solution of

$$dX_t(\theta) = (b(X_t(\theta)) + \sigma(X_t(\theta))\theta)\, dt + \sigma(X_t(\theta))dW_t, \tag{4.3}$$

then

$$\mathbb{E}[\psi(X_T)] = \mathbb{E}\left[\psi(X_T(\theta))\, \mathrm{e}^{-\theta\cdot W_T - \frac{1}{2}|\theta|^2 T}\right], \quad \forall \theta \in \mathbb{R}^q. \tag{4.4}$$

From now on, we assume that

$$\mathbb{P}(\psi(X_T) \neq 0) > 0 \quad \text{and} \quad \forall \theta \in \mathbb{R}^q,\ \mathbb{E}[\psi(X_T)^2\, \mathrm{e}^{-\theta \cdot W_T}] < +\infty. \tag{4.5}$$

For $\alpha > 0$, we introduce the space

$$\mathcal{H}_\alpha = \Big\{ \psi : \mathbb{R}^d \to \mathbb{R} \text{ s.t. } \exists\, c > 0, \beta \geq 1,\ \forall x \in \mathbb{R}^d,\ |\psi(x)| \leq c(1 + |x|^\beta) \\ \text{and } \forall x, y \in \mathbb{R}^d,\ |\psi(x) - \psi(y)| \leq c(1 + (|x|^\beta \wedge |y|^\beta))|x - y|^\alpha \Big\}. \tag{4.6}$$

For any function $\psi \in \mathcal{H}_\alpha$, (4.5) implies that $\sup_n \mathbb{E}[\psi(X_T^n)^2\, \mathrm{e}^{-\theta \cdot W_T}] < +\infty$. We also introduce the continuous time Euler approximation $X^n(\theta)$ of the process $X(\theta)$. It is natural to choose the value of θ minimizing $\mathrm{Var}\left(\psi(X_T(\theta))\, \mathrm{e}^{-\theta \cdot W_T - \frac{1}{2}|\theta|^2 T}\right)$, we set

$$\theta_\star = \underset{\theta \in \mathbb{R}^q}{\operatorname{argmin}}\, v(\theta) \quad \text{with} \quad v(\theta) = \mathbb{E}\left[\psi(X_T)^2 e^{-\theta \cdot W_T + \frac{1}{2}|\theta|^2 T}\right]. \tag{4.7}$$

From a practical point of view, the quantity $v(\theta)$ is not explicit so we use the Euler scheme to discretize $X(\theta)$ and approximate $\theta_\star$ by

$$\theta_n = \underset{\theta \in \mathbb{R}^q}{\operatorname{argmin}}\, v_n(\theta) \quad \text{with} \quad v_n(\theta) = \mathbb{E}\left[\psi(X_T^n)^2 e^{-\theta \cdot W_T + \frac{1}{2}|\theta|^2 T}\right]. \tag{4.8}$$

The functions v and v_n are evaluated using the original diffusion X and their dependency on the shift is restricted to the exponential weight, which will play a key role to obtain regularity properties for v and v_n; we already used this methodology in Chapter 3. Since the expectation is usually not tractable, we aim at using a sample average approximation procedure to approximate θ_n

$$\theta_{n,N} = \underset{\theta \in \mathbb{R}^q}{\operatorname{argmin}}\, v_{n,N}(\theta) \quad \text{with} \quad v_{n,N}(\theta) = \frac{1}{N} \sum_{i=1}^{N} \left(\psi(X_{T,i}^n)^2 e^{-\theta \cdot W_{T,i} + \frac{1}{2}|\theta|^2 T}\right), \tag{4.9}$$

where $(X_{T,i}^n, W_{T,i})_{1 \leq i \leq N}$ are i.i.d. samples according to the law of (X_T^n, W_T). The existence and uniqueness of $\theta_\star$, θ_n and $\theta_{n,N}$ are ensured by the following Lemma, which can be easily deduced from Proposition 3.1.1 or Proposition 3.2.2.

Lemma 4.2.1 *Under Condition* (4.5)*, the functions v, v_n and $v_{n,N}$ are infinitely continuously differentiable for all $n, N \geq 1$ and for all multi-index $r \in \mathbb{N}^q$, we have*

$$\partial_\theta^r v(\theta) = \mathbb{E}\left[\partial_\theta^r \left(\psi(X_T)^2 e^{-\theta \cdot W_T + \frac{1}{2}|\theta|^2 T}\right)\right];\ \partial_\theta^r v_n(\theta) = \mathbb{E}\left[\partial_\theta^r \left(\psi(X_T^n)^2 e^{-\theta \cdot W_T + \frac{1}{2}|\theta|^2 T}\right)\right].$$

Moreover, the functions v, v_n and $v_{n,N}$ are strongly convex for any $n \geq 1$, and any $N \geq 1$ such that $\psi(X_{T,i}^n) > 0$ for some $i \leq N$.

4.2.1 Convergence of the optimal importance sampling parameters

From [13, Theorem 2.2], we have the following result.

Theorem 4.2.2 *Let ψ satisfy Condition* (4.5) *and belong to $\mathcal{H}_\alpha$ for some $\alpha > 0$. Then, $\theta_n \to \theta_\star$ a.s. when $n \to \infty$.*

From now on, we let N depend on n so that $N = N_n$ is an increasing function of n and tends to infinity with n.

Proposition 4.2.3 *Assume that $\psi \in \mathcal{H}_\alpha$ for some $\alpha > 0$. Then, for all $K > 0$, a.s.*

$$\sup_{|\theta|\leq K} |v_{n,N_n}(\theta) - v(\theta)| \xrightarrow[n\to+\infty]{} 0; \quad \sup_{|\theta|\leq K} |\nabla v_{n,N_n}(\theta) - \nabla v(\theta)| \xrightarrow[n\to+\infty]{} 0.$$

Proof. As the proof of the two results are very similar, we only focus on the uniform convergence for v_{n,N_n}. To do so, we will apply Proposition 7.1.3. Amongst the assumptions required to apply this result, only ($\mathcal{H}7.4$) needs some explanation. If we fix $\delta > 0$ and $\theta \in \mathbb{R}^d$, then we have by the Cauchy Schwartz inequality

$$\sup_n \mathbb{E}\left[\psi(X_T^n)^2 \sup_{|\theta'-\theta|\leq\delta} \left|e^{-\theta'\cdot W_T+\frac{1}{2}|\theta'|^2T} - e^{-\theta\cdot W_T+\frac{1}{2}|\theta|^2T}\right|\right]^2 \leq$$
$$\sup_n \mathbb{E}\left[\psi(X_T^n)^4\right] \mathbb{E}\left[\sup_{|\theta-\theta|\leq\delta} \left|e^{-\theta'\cdot W_T+\frac{1}{2}|\theta'|^2T} - e^{-\theta\cdot W_T+\frac{1}{2}|\theta|^2T}\right|^2\right].$$

Using the elementary algebraic inequality $|e^x - e^y| \leq |x-y|\,(\mathrm{e}^x + \mathrm{e}^y)$, we easily deduce that the quantity $\mathbb{E}\left[\sup_{|\theta'-\theta|\leq\delta} \left|e^{-\theta'\cdot W_T+\frac{1}{2}|\theta'|^2T} - e^{-\theta\cdot W_T+\frac{1}{2}|\theta|^2T}\right|^2\right]$ can be made arbitrarily small. Finally, we apply Remark 7.1.4 to show that Assumption ($\mathcal{H}7.4$) holds. ■

Theorem 4.2.4 *Assume that $\psi \in \mathcal{H}_\alpha$ for some $\alpha > 0$. Then, $\theta_{n,N_n} \xrightarrow[n\to+\infty]{a.s.} \theta^*$, $v_{n,N_n}(\theta_{n,N_n}) \xrightarrow[n\to+\infty]{a.s.} v(\theta_\star)$ and $\sqrt{N_n}(\theta_{n,N_n} - \theta^*) \xrightarrow[n\to+\infty]{law} N(0,\Gamma)$ where*

$$\Gamma = [\nabla^2 v(\theta_\star)]^{-1} \operatorname{Cov}\left[(T\theta_\star - W_t)\psi(X_T)^2\,\mathrm{e}^{-\theta_\star\cdot W_T+\frac{1}{2}|\theta_\star|^2T}\right][\nabla^2 v(\theta_\star)]^{-1}.$$

Sketch of the proof. We already know from Proposition 4.2.3 that a.s. v_{n,N_n} converges local uniformly to v. We can reproduce the proof of Proposition 3.2.5 to obtain the a.s. convergence of $v_{n,N_n}(\theta_{n,N_n})$. Moreover, for all $K > 0$

$$\sup_{|\theta|\leq K} \left|\partial_{\theta^{(j)}}(\psi(X_T)^2\,\mathrm{e}^{-\theta\cdot W_T+\frac{1}{2}|\theta|^2T})\right|$$
$$\leq \mathrm{e}^{K^2T/2}\,\psi(X_T)^2\left(K + (\mathrm{e}^{KW_t^{(j)}} + \mathrm{e}^{-KW_t^{(j)}})\right)\prod_{i=1}^q (\mathrm{e}^{KW_t^{(i)}} + \mathrm{e}^{-KW_t^{(i)}}).$$

The r.h.s is integrable by (4.5). Hence, $\mathbb{E}\left[\sup_{|\theta|\leq K}\left|\nabla_\theta(\psi(X_T)^2\,\mathrm{e}^{-\theta\cdot W_T+\frac{1}{2}|\theta|^2T})\right|\right] < +\infty$. Similarly, one can prove that $\mathbb{E}\left[\sup_{|\theta|\leq K}\left|\nabla^2_\theta(\psi(X_T)^2\,\mathrm{e}^{-\theta\cdot W_T+\frac{1}{2}|\theta|^2T})\right|\right] < +\infty$. Then, to prove the central limit theorem governing the convergence of θ_{n,N_n} to $\theta_\star$, we reproduce the proof of Rubinstein and Shapiro [83, Theorem A2, pp. 74], which is mainly based on the a.s. locally uniform convergence of $\nabla v_{n,N_n}$ and on the asymptotic normality of $v_{n,N_n}(\theta_\star)$ ensuing from the Lindeberg–Feller theorem for triangular array of random variables (see [18]). ■

4.2.2 Strong law of large numbers and central limit theorem

Based on the methodology developed in Section 3.3, we define a Monte Carlo estimator of $\mathbb{E}[\psi(X_T)]$ based on Equation (4.4). We introduce the σ-algebra $\mathcal{G}$ generated by the samples $(W_i)_{i\geq 1}$ used to compute θ_n and θ_{n,N_n}.

Let $(\tilde{W}_i)_i$ be i.i.d. samples according to the law of W but independent of $\mathcal{G}$. We introduce the i.i.d. samples $(\tilde{X}_i(\theta))_i$ following the law of $X(\theta)$ such that for each i, $\tilde{X}_i(\theta)$ is the solution of the SDE (4.3) driven by $\tilde{W}_i$. We introduce $(\tilde{\mathcal{G}}_k)_{k>0}$ the filtration defined by $\tilde{\mathcal{G}}_k = \sigma(\tilde{W}_i, 1 \leq i \leq k)$ and $\mathcal{G}_k^\sharp = \mathcal{G} \vee \tilde{\mathcal{G}}_k$. For each $i > 0$, we also introduce the Euler discretization $\tilde{X}_i^n(\theta)$ of $\tilde{X}_i(\theta)$. Based on these new sets of samples, we introduce the Monte Carlo estimator

$$M_{n,N_n}(\theta) = \frac{1}{N_n} \sum_{i=1}^{N_n} g(\theta, \tilde{X}_{T,i}^n(\theta), \tilde{W}_{T,i})$$

where the function $g : \mathbb{R}^q \times \mathbb{R}^d \times \mathbb{R}^q \to \mathbb{R}$ is defined by

$$g(\theta, x, y) = \psi(x)\, \mathrm{e}^{-\theta \cdot y - \frac{1}{2}|\theta|^2 T}\,. \tag{4.10}$$

Theorem 4.2.5 *Assume that $\psi \in \mathcal{H}_\alpha$ for some $\alpha > 0$. Then, $M_{n,N_n}(\theta_{n,N_n}) \longrightarrow \mathbb{E}[\psi(X_T)]$ a.s. when $n \to +\infty$.*

Proof. Using the conditional independence of the samples $(\tilde{X}_i^n(\theta_{n,N_n}), \tilde{W}_i)_i$, we have

$$\mathbb{E}[g(\theta_{n,N_n}, \tilde{X}_{T,i}^n(\theta_{n,N_n}), \tilde{W}_{T,i})|\mathcal{G}] = \mathbb{E}[\psi(X_T^n)] = e_n \quad \text{for all } i > 0.$$

Let $\mathcal{V} \subset \mathbb{R}^q$ be a compact neighbourhood of $\theta_\star$. We define the sequence

$$Y_{i,n} = \Big(g(\theta_{n,N_n}, \tilde{X}_{T,i}^n(\theta_{n,N_n}), \tilde{W}_{T,i}) - e_n \Big)\, \mathbf{1}_{\theta_{n,N_n} \in \mathcal{V}}$$

and its empirical average $\overline{Y}_{m,n} = \frac{1}{m}\sum_{i=1}^m Y_{i,n}$ for all $m > 0$. It is obvious that $\mathbb{E}[Y_{i,n}] = 0$ and using the conditional independence $\mathbb{E}[\big|\overline{Y}_{m,n}\big|^2] = \frac{1}{m}\mathbb{E}[|Y_{1,n}|^2]$.

$$\begin{aligned}
\mathbb{E}[|Y_{1,n}|^2] &\leq \mathbb{E}\left[\mathbb{E}\left[|g(\theta_{n,N_n}, \tilde{X}_{T,i}^n(\theta_{n,N_n}), \tilde{W}_{T,i}) - e_n|^2 \Big| \mathcal{G}\right] \mathbf{1}_{\theta_{n,N_n} \in \mathcal{V}}\right] \\
&\leq \mathbb{E}\left[v_n(\theta_{n,N_n}) \mathbf{1}_{\theta_{n,N_n} \in \mathcal{V}}\right] \leq \sup_{\theta \in \mathcal{V}} v_n(\theta).
\end{aligned}$$

We know that v_n is convex and converges point-wise to v, which is also convex and continuous. Hence, v_n converges locally uniformly to v, which implies that for all compact sets $K \subset \mathbb{R}^q$, $\lim_{n\to+\infty} \sup_{\theta\in K} v_n(\theta) = \sup_{\theta\in K} v(\theta)$. Hence, $\sup_n \sup_{\theta\in\mathcal{V}} v_n(\theta) < +\infty$. Applying Proposition 7.1.1 proves that $\overline{Y}_{N_n,n} \xrightarrow[n\to+\infty]{a.s.} 0$. As θ_{n,N_n} converges a.s. to $\theta^* \in K$, this also implies that $\lim_{n\to+\infty} M_{n,N_n} = \mathbb{E}[\psi(X_T)]$ a.s. ■

Theorem 4.2.6 *Under the assumptions of Theorem 4.2.5 and if* (4.2) *holds, we have*

$$\sqrt{N_n}(M_{n,N_n}(\theta_{n,N_n}) - \mathbb{E}[\psi(X_T)]) \xrightarrow[n\to\infty]{\mathcal{D}} \mathcal{N}(C_\psi(T,\gamma), v(\theta_\star) - \mathbb{E}[\psi(X_T)]^2).$$

Sketch of the proof. We can write the left hand side of the convergence result by introducing $M_{n,N_n}(\theta_\star)$

$$\sqrt{N_n}(M_{n,N_n} - \mathbb{E}[\psi(X_T)]) = \sqrt{N_n}(M_{n,N_n}(\theta_{n,N_n}) - M_n(\theta_\star)) + \sqrt{N_n}(M_{n,N_n}(\theta_\star) - \mathbb{E}[\psi(X_T)])$$

The convergence of the last term on the r.h.s $\sqrt{N_n}(M_{n,N_n}(\theta_\star) - \mathbb{E}[\psi(X_T)])$ is governed by the central limit theorem for Euler Monte Carlo, which yields the announced limit (see [36]). It remains to prove that $\sqrt{N_n}(M_{n,N_n}(\theta_{n,N_n}) - M_{n,N_n}(\theta_\star))$ converges to zero in probability, which is achieved by closely following the methodology used in the proof of Theorem 3.3.4. ■

4.3 The importance sampling multilevel estimator

4.3.1 The general setting

The multilevel idea. Multilevel Monte Carlo estimators smartly combine together discretization schemes on nested time grids. Let $m \in \mathbb{N}$ such that $m \geq 2$ be the number of ticks of the coarsest time grid. Let $L \in \mathbb{N}^*$, we consider the set of nested time grids with m^ℓ ticks for $\ell = 1, \cdots, L$. Write

$$\mathbb{E}\left[\psi(X_T^{m^L})\right] = \mathbb{E}[\psi(X_T^{m^0})] + \sum_{\ell=1}^{L} \mathbb{E}\left[\psi(X_T^{m^\ell}) - \psi(X_T^{m^{\ell-1}})\right]. \tag{4.11}$$

Each expectation is approximated by a Monte Carlo method independent of all the others, which leads to the following estimator

$$Q_L = \frac{1}{N_0}\sum_{k=1}^{N_0} \psi(\tilde{X}_{T,0,k}^{m^0}) + \sum_{\ell=1}^{L} \frac{1}{N_\ell}\sum_{k=1}^{N_\ell} \left(\psi(\tilde{X}_{T,\ell,k}^{m^\ell}) - \psi(\tilde{X}_{T,\ell,k}^{m^{\ell-1}})\right) \tag{4.12}$$

where for each $\ell \in \{1, \cdots, L\}$, N_ℓ is the number of samples used to build a Monte Carlo estimator of the expectation involved in level ℓ and the random variables $\tilde{X}_{T,\ell,k}^{m^\ell}$ (resp. $\tilde{X}_{T,\ell,k}^{m^{\ell-1}}$) are the terminal values of the Euler schemes of X with m^ℓ (resp. $m^{\ell-1}$)) time steps built using the same Brownian paths. The blocks of random variables used in two different levels are independent.

Its is known from Ben Alaya et al. [13] that for properly chosen N_ℓ's, the estimator defined by (4.12) satisfies a central limit theorem with limiting variance given by $\mathrm{Var}(\nabla\psi(X_T)\cdot U_T)$ where the process U is the limit of $\sqrt{n}(X^n - X)$ (see [52]).

A smart multilevel importance sampling estimator. We define

$$\forall\, x \in \mathbb{R}^q, \quad \mathcal{E}^-(x,\theta) = \mathrm{e}^{-\theta\cdot x - \frac{1}{2}|x|^2 T}; \quad \mathcal{E}^+(x,\theta) = \mathrm{e}^{-\theta\cdot x + \frac{1}{2}|x|^2 T}.$$

One way of introducing importance sampling is to apply the multilevel approach to $\psi(X_T(\theta))\,\mathrm{e}^{-\theta\cdot W_T - |\theta|^2 T/2}$. Then, (4.11) becomes

$$\mathbb{E}\left[\psi(X_T^{m^L}(\theta))\mathcal{E}^-(W_T,\theta)\right] =$$
$$\mathbb{E}[\psi(X_T^{m^0}(\theta))\mathcal{E}^-(W_T,\theta)] + \sum_{\ell=1}^{L} \mathbb{E}\left[\left\{\psi(X_T^{m^\ell}(\theta)) - \psi(X_T^{m^{\ell-1}}(\theta))\right\}\mathcal{E}^-(W_T,\theta)\right].$$

Then, the optimal choice for θ would be $\operatorname{argmin}_\theta \mathbb{E}[(\nabla\psi(X_T)\cdot U_T)^2\mathcal{E}^+(W,\theta)]$, which in turn needs some discretization scheme to be approximated. This leads to lots of extra computations and implementations.

A *smart* way to circumvent this practical drawback is to apply importance sampling to each level with its own parameter θ. We define the multilevel importance sampling estimator by

$$Q_L(\lambda_0,\ldots,\lambda_L) = \frac{1}{N_0}\sum_{k=1}^{N_0} \psi(\tilde{X}_{T,0,k}^{m^0}(\lambda_0))\mathcal{E}^-(\tilde{W}_{0,k},\lambda_0)$$
$$+ \sum_{\ell=1}^{L} \frac{1}{N_\ell}\sum_{k=1}^{N_\ell} \left(\psi(\tilde{X}_{T,\ell,k}^{m^\ell}(\lambda_\ell)) - \psi(\tilde{X}_{T,\ell,k}^{m^{\ell-1}}(\lambda_\ell))\right)\mathcal{E}^-(\tilde{W}_{\ell,k},\lambda_\ell)$$

for any $\Lambda_L = (\lambda_0, \ldots, \lambda_L) \in (\mathbb{R}^q)^L$. For every ℓ, The random variables $(\tilde{W}_{\ell,k})_{1\leq k\leq N_\ell}$ are independent and are distributed according to the Brownian law. We assume that for $\ell, \ell' \in \{1, \cdots, L\}$, with $\ell \neq \ell'$, the blocks $(\tilde{W}_{\ell,k})_{1\leq k\leq N_\ell}$ and $(\tilde{W}_{\ell',k})_{1\leq k\leq N_{\ell'}}$ are independent. The variables $\tilde{X}^{m^\ell}_{T,\ell,k}$ and $\tilde{X}^{m^{\ell-1}}_{T,\ell,k}$ are built using the Brownian $\tilde{W}_{\ell,k}$. The variance of the multilevel estimator is given by

$$\mathrm{Var}[Q_L] = N_0^{-1}\,\mathrm{Var}[\psi(X_T^{m^0}(\lambda_0))\mathcal{E}^-(W,\lambda_0)] + \sum_{\ell=1}^{L} N_\ell^{-1}\frac{(m-1)T}{m^\ell}\sigma_\ell^2(\lambda_\ell)$$

where

$$\sigma_\ell^2(\lambda) = \frac{m^\ell}{(m-1)T}\,\mathrm{Var}\left[\left\{\psi(X_T^{m^\ell}(\lambda)) - \psi(X_T^{m^{\ell-1}}(\lambda))\right\}\mathcal{E}^-(W,\lambda)\right].$$

The variance of level ℓ can be rewritten as $\sigma_\ell^2(\lambda) = v_\ell(\lambda) - \frac{m^\ell}{(m-1)T}\mathbb{E}\left[\psi(X_T^{m^\ell}) - \psi(X_T^{m^{\ell-1}})\right]^2$ with

$$v_\ell(\lambda) = \frac{m^\ell}{(m-1)T}\mathbb{E}\left[\left|\psi(X_T^{m^\ell}) - \psi(X_T^{m^{\ell-1}})\right|^2\mathcal{E}^+(W,\lambda)\right]. \tag{4.13}$$

In the spirit of Chapter 3, we define the *Sample Average Approximation* of v_ℓ by

$$v_{\ell,N'_\ell}(\lambda) = \frac{1}{N'_\ell}\sum_{k=1}^{N'_\ell}\frac{m^\ell}{(m-1)T}\left|\psi(X^{m^\ell}_{T,\ell,k}) - \psi(X^{m^{\ell-1}}_{T,\ell,k})\right|^2\mathcal{E}^+(W_{\ell,k},\lambda) \tag{4.14}$$

where the variables W_k are i.i.d. according to the Brownian law on $[0,T]$ and are independent of the $\tilde{W}_k$'s. Based on these new Brownian paths, we introduce the random variables $X^{m^\ell}_{T,k}$, defined in the same way as the tilde quantities but independent of them. Hence, the estimators v_{ℓ,N'_ℓ} for $\ell = 1,\ldots,L$ are independent of $Q_L(\lambda_0,\ldots,\lambda_L)$. The variance of level ℓ only depends on λ_ℓ, so minimizing the total variance $\mathrm{Var}[Q_L]$ is achieved by independently minimizing the variance of each level.

Note that the number N'_ℓ of samples used to build a Monte Carlo approximation of v_ℓ may differ from the number N_ℓ of samples used in the computation of the level ℓ of Q_L. This point will be discussed in details in the numerical section. For the moment, we just require N'_ℓ to go to infinity with ℓ.

By Lemma 4.2.1, the functions v_ℓ and v_{ℓ,N'_ℓ} are strongly convex and infinitely differentiable. Hence, we can define

$$\widehat{\lambda}_\ell = \arg\min_{\lambda\in\mathbb{R}^q} v_{\ell,N'_\ell}(\lambda).$$

Theorem 4.3.1 *Assume b and σ are C^1 with bounded derivatives, $\psi \in \mathcal{H}_\alpha$ for some $\alpha \geq 1$, ψ is C^1 and $\nabla\psi$ has polynomial growth. Then, the sequence of random functions v_{ℓ,N'_ℓ} converges a.s. locally uniformly to the strongly convex function $v : \mathbb{R}^q \to \mathbb{R}$ defined by*

$$v(\lambda) = \mathbb{E}\left[(\nabla\psi(X_T)\cdot U_T)^2\,\mathcal{E}^+(W,\lambda)\right]. \tag{4.15}$$

Moreover, $\widehat{\lambda}_\ell$ converges a.s. to $\lambda_\star = \arg\min_\lambda v(\lambda)$, when $\ell \to +\infty$.

4.3.2 Strong law of large numbers and central limit theorem

In this section, we state our two main results dealing with the convergence of $Q_L(\hat{\lambda}_0, \cdots, \hat{\lambda}_L)$.

We assume that the sample size N_ℓ has the following form

$$N_{\ell,L} = \frac{\rho(L)}{m^\ell a_\ell} \sum_{k=1}^{L} a_k, \quad \ell \in \{0, \cdots, L\} \tag{4.16}$$

for some increasing function $\rho : \mathbb{N} \to \mathbb{R}$ and sequence $(a_\ell)_{\ell \in \mathbb{N}}$ of positive real numbers such that $\lim_{L\to\infty} \sum_{\ell=1}^{L} a_\ell = \infty$. We recall that for any sequence $(x_\ell)_{\ell \geq 1}$ converging to some limit $x \in \mathbb{R}$,

$$\lim_{L\to+\infty} \frac{\sum_{\ell=1}^{L} a_\ell x_\ell}{\sum_{\ell=1}^{L} a_\ell} = x.$$

Theorem 4.3.2 *Assume that* $\sup_L \sup_\ell \frac{L^2 a_\ell}{\rho(L)\sum_{k=1}^{L} a_k} < +\infty$. *Then, under the assumptions of Theorem 4.3.1,* $Q_L(\widehat{\lambda}_0, \ldots, \widehat{\lambda}_L) \longrightarrow \mathbb{E}[\psi(X_T)]$ *a.s. when* $L \to +\infty$.

For the choice $a_\ell = 1$ for all ℓ, the condition on ρ reduces to $\sup_L \frac{L}{\rho(L)} < +\infty$. This result is proved by applying Proposition 7.1.1 to each level. As the computations quickly become tedious, we refer the reader to [L-1] for the complete proof of this result.

Theorem 4.3.3 *Suppose that the assumptions of Theorem 4.3.1 hold and that* (4.2) *is satisfied. Then, for* $N_{\ell,L}$ *given by* (4.16) *with* $\rho(L) = m^{2\gamma L}(m-1)T$ *and the sequence* $(a_\ell)_\ell$ *satisfying*

$$\lim_{L\to\infty} \frac{1}{\left(\sum_{\ell=1}^{L} a_\ell\right)^{p/2}} \sum_{\ell=1}^{L} a_\ell^{p/2} = 0, \quad \textit{for } p > 2. \tag{4.17}$$

We have

$$m^{\gamma L}(Q_L(\widehat{\lambda}_0, \ldots, \widehat{\lambda}_L) - \mathbb{E}[\psi(X_T)]) \xrightarrow[L\to+\infty]{\mathcal{D}} \mathcal{N}(C_\psi(T,\gamma), v(\lambda^*))$$

where the function v *is defined by* (4.15).

The convergence rate does not depend on the numbers of samples N'_ℓ provided that they tend to infinity with ℓ. This convergence basically ensues from the central limit theorem for martingale arrays.

4.4 Numerical experiments

4.4.1 Practical implementation

Our approach cleverly mixes the famous multilevel Monte Carlo technique with importance sampling to reduce the variance. A classical approach would have been to consider the multilevel approximation of $\mathbb{E}\left[\psi(X_T(\theta))\,\mathrm{e}^{-\theta\cdot W_T - \frac{1}{2}|\theta|^2 T}\right]$ while choosing the value of θ which minimizes the variance of the central limit theorem for multilevel Monte Carlo (see [12]). This asymptotic variances involves both $\nabla\psi$ and the process U. Hence, a classical approach to importance sampling for multilevel Monte Carlo would require extra knowledge than the function ψ and the underlying process X, thus precluding any kind of automation.

We have chosen a completely different approach allowing for one importance sampling parameter per level, which enables us to treat each level independently of the others. In each level, we use a sample average approximation as in Chapter 3 to compute the optimal importance sampling parameter defined as the one minimizing the variance of the current level. From Theorem 4.3.3, we know that

this approach is optimal in the sense that our multilevel estimator $Q_L(\hat{\lambda}_0, \dots, \hat{\lambda}_L)$ satisfies a central limit theorem with a limiting variance given by $\inf v$ where v defined by (4.15) is the variance of the standard multilevel Monte Carlo estimator. We managed to provide an algorithm reaching the optimal limiting variance without computing $\nabla\psi$ nor the process U, hence our approach can be made fully automatic. Our overall algorithm is described in Algorithm 4.4.1.

1 Generate $X_{T,0,1}^{m^0}, \dots, X_{T,0,N_0'}^{m^0}$ i.i.d. samples following the law of $X_T^{m^0}$ independently of the other blocks.
2 Compute the minimizer $\hat{\lambda}_0$ of $u_{0,N_0'}$ by solving $\nabla u_{0,N_0'}(\hat{\lambda}_0) = 0$.
3 **for** $\ell = 1 : L$ **do**
4 Generate $(X_{T,\ell,1}^{m^\ell}, X_{T,\ell,1}^{m^{\ell-1}}), \dots, (X_{T,\ell,N_\ell'}^{m^\ell}, X_{T,\ell,N_\ell'}^{m^{\ell-1}})$ i.i.d. samples following the law of $(X_T^{m^\ell}, X_T^{m^{\ell-1}})$ independently of the other blocks.
5 Compute the minimizer $\hat{\lambda}_\ell$ of of $u_{\ell,N_\ell'}$ by solving $\nabla u_{\ell,N_\ell'}(\hat{\lambda}_\ell) = 0$.
6 **end**
7 Conditionally on $\hat{\lambda}_0$, generate $\tilde{X}_{T,0,1}^{m^0}(\hat{\lambda}_0), \dots, \tilde{X}_{T,0,N_0}^{m^0}(\hat{\lambda}_0)$ i.i.d. samples with the law of $X_T^{m^0}(\hat{\lambda}_0)$ independently of the other blocks. The tilde and non tilde quantities are conditionally independent.
8 **for** $\ell = 1 : L$ **do**
9 Conditionally on $\hat{\lambda}_\ell$, generate $(\tilde{X}_{T,\ell,1}^{m^\ell}(\hat{\lambda}_\ell), \tilde{X}_{T,\ell,1}^{m^{\ell-1}}(\hat{\lambda}_\ell)), \dots, (\tilde{X}_{T,\ell,N_\ell}^{m^\ell}(\hat{\lambda}_\ell), \tilde{X}_{T,\ell,N_\ell}^{m^{\ell-1}}(\hat{\lambda}_\ell))$ i.i.d. samples with the law of $(X_T^{m^\ell}(\hat{\lambda}_\ell), X_T^{m^{\ell-1}}(\hat{\lambda}_\ell))$ independently of the other blocks. The tilde and non tilde quantities are conditionally independent.
10 **end**
11 Compute the multilevel importance sampling estimator

$$Q_L(\hat{\lambda}_0, \dots, \hat{\lambda}_L) = \frac{1}{N_0} \sum_{k=1}^{N_0} \psi(\tilde{X}_{T,0,k}^{m^0}(\hat{\lambda}_0)) \mathcal{E}^-(\tilde{W}_{0,k}, \hat{\lambda}_0) + \sum_{\ell=1}^{L} \frac{1}{N_\ell} \sum_{k=1}^{N_\ell} \left(\psi(\tilde{X}_{T,\ell,k}^{m^\ell}(\hat{\lambda}_\ell)) - \psi(\tilde{X}_{T,\ell,k}^{m^{\ell-1}}(\hat{\lambda}_\ell)) \right) \mathcal{E}^-(\tilde{W}_{\ell,k}, \hat{\lambda}_\ell).$$

Algorithm 4.4.1: Multilevel Importance Sampling (MLIS)

The minimization step (items 2 and 4 in Algorithm 4.4.1) is performed using a Newton algorithm. Unlike what happens in a classical Monte Carlo method in which a new sample is drawn at each iteration, here all the samples must be stored since the same random variables are used in all the iterations of the Newton procedure. This feature is specific to the optimisation step and may make the algorithm highly memory demanding as soon as the numbers N_ℓ' become large. As the parameter λ is not involved in the function ψ, all the quantities $\psi(X_{T,\ell,k}^{m^\ell}) - \psi(X_{T,\ell,k}^{m^{\ell-1}})$ for $k = 1, \dots, N_\ell$ can be precomputed before starting the minimization algorithm, which enables us to save a lot of computational time. The efficiency of Newton's algorithm very much depends on the convexity of the v_{ℓ,N_ℓ} functions.

Complexity analysis. In this paragraph, we focus on the impact of the number of levels L on the overall computational time of our algorithm. The computational cost of the standard multilevel esti-

mator is proportional to

$$C_{ML} = \sum_{\ell=0}^{L} N_\ell m^\ell = m^{2L+1} L^2.$$

The global cost of our algorithm writes as the sum of the cost of the computation of the $(\hat{\lambda}_\ell)_\ell$ and of the standard multilevel estimator

$$C_{MLIS} = \sum_{\ell=0}^{L} N'_\ell (m^\ell + 3K_\ell) + \sum_{\ell=0}^{L} N_\ell m^\ell$$

where K_ℓ is the number of iterations of Newton's algorithm to approximate $\hat{\lambda}_\ell$ and the factor 3 corresponds to the fact that building $\nabla u_{\ell,N'_\ell}$ and $\nabla^2 u_{\ell,N'_\ell}$ basically boils down to three Monte Carlo summations. In practice, $K_\ell \leq 5$ as the problem is strongly convex. Because the same random variables are used at each iteration of the optimisation step, they must be stored, which makes the memory footprint of our algorithm proportional to N'_ℓ.

So, if we choose $N'_\ell = \frac{N_\ell m^\ell}{m^\ell + 15}$, the total cost of our MLIS algorithm should be roughly twice the cost of the standard multilevel estimator. This choice of N'_ℓ reduces the number of samples used to approximate the variance of the first levels compared to using directly N_ℓ. However, when L increases, N'_ℓ can become extremely large for small values of ℓ which leads to an even larger memory footprint (see Section 4.4.1). To avoid breaking the scalability of the algorithm, the values of N'_ℓ have to be kept reasonable depending on the amount of memory available on the computer. For instance, enforcing $N'_\ell \leq 500000$ is reasonable on a computer with $8Gb$ of RAM. Anyway, it is crystal clear that a fairly good approximation of the variance v_ℓ is enough and running for an ultimately accurate estimator would lead to a tremendous waste of computational time. Monitoring the convergence of v_{ℓ,N'_ℓ} would really help choosing sensible values for N'_ℓ.

4.4.2 Experimental settings

We compare four methods in terms of their root mean squared error (RMSE): the crude Monte Carlo method (MC), the adaptive Monte Carlo method proposed in Chapter 3 (MC+IS), the Multilevel Monte Carlo method (ML) and our Importance Sampling Multilevel Monte Carlo estimator (ML+IS). We recall that the RMSE is defined by $RMSE = \sqrt{\text{Bias}^2 + \text{Variance}}$. In the computation of the bias, the true value is replaced by its multilevel Monte Carlo estimator with $L = 9$ levels, which yields a very accurate approximation. Not to mention, the CPU times showed on the graphs take into account both the time for the search of the optimal parameter and the time for the second stage Monte Carlo, be it multilevel or not.

4.4.3 Multidimensional Dupire's framework

We consider a $d-$dimensional local volatility model, in which the dynamics, under the risk neutral measure, of each asset S^i is supposed to be given by

$$dS^i_t = S^i_t(r\,dt + \sigma(t, S^i_t)dW^i_t), \qquad S_0 = (S^1_0, \ldots, S^d_0)$$

where $W = (W^1, \ldots, W^d)$, each component W^i being a standard Brownian motion with values in $\mathbb{R}$. For the numerical experiments, the covariance structure of W will be assumed to be given by $\langle W^i, W^j \rangle_t = \rho t \mathbf{1}_{i \neq j} + t \mathbf{1}_{i=j}$. We suppose that $\rho \in (-\frac{1}{d-1}, 1)$, which ensures that the matrix

$C = (\rho \mathbf{1}_{i \neq j} + \mathbf{1}_{i=j})_{1 \leq i,j \leq d}$ is positive definite. The maturity time and the interest rate are respectively denoted by $T > 0$ and $r > 0$. The local volatility function σ we have chosen is of the form

$$\sigma(t,x) = 0.6(1.2 - \mathrm{e}^{-0.1t}\,\mathrm{e}^{-0.001(x\,\mathrm{e}^{rt}-s)^2})\,\mathrm{e}^{-0.05\sqrt{t}}, \tag{4.18}$$

with $s > 0$. We know that there exists a duality between the variables (t,x) and (T,K) in Dupire's framework. Hence for formula (4.18) to make sense, one should choose s equal to the spot price of the underlying asset so that the bottom of the smile is located at the forward money. We refer to Figure 4.4.1 to have an overview of the smile.

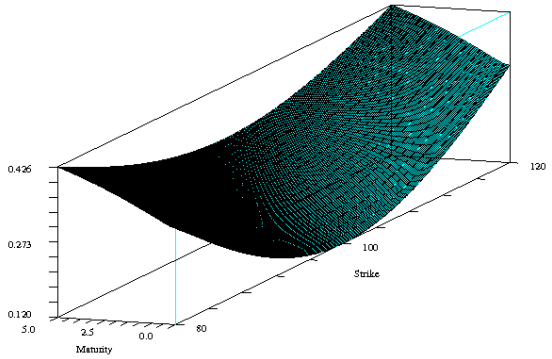


Figure 4.4.1: Local volatility function

Basket option We consider options with payoffs of the form $(\sum_{i=1}^{d} \omega^i S_T^i - K)_+$ where $(\omega^1, \ldots, \omega^d)$ is a vector of algebraic weights. The strike value K can be taken negative to deal with Put like options. With no surprise, we can see on Figure 4.4.2 that multilevel estimators always outperform their classical Monte Carlo counterparts. The comparison for very little accurate estimators may be meaningless as it is pretty difficult to reliably measure short execution times and the empirical variance of the estimator is in this case even less accurate than the estimator itself. Note that the points on the extreme right hand side are obtained for multilevel estimators with $L = 2$, respectively for Monte Carlo estimators with 256 samples. For RMSE between 0.1 and 0.005, our MLIS estimator is 10 times faster than the standard ML estimator. When a very high accuracy is required, namely when RMSE is smaller than 0.001, the MLIS estimator remains between 3 and 4 times faster than the standard multilevel estimator, which is already a great achievement since for this level of accuracy, the ML estimator may need several dozens of minutes to yield its result.

4.4.4 Multidimensional Heston model

The multidimensional Heston model can be easily written by specifying on the one hand that each asset follows a 1-D Heston model and on the other hand the correlation structure between the involved Brownian motions. The asset price process $S = (S^1, \ldots, S^d)$ and the volatility process $\sigma = (\sigma^1, \ldots, \sigma^d)$ solve

$$\begin{aligned} dS_t^i &= rS_t^i dt + \sqrt{\sigma_t^i} S_t^i dB_t^i \\ d\sigma_t^i &= \kappa^i(a^i - \sigma_t^i)dt + \nu_t^i \sqrt{\sigma_t^i}(\gamma^i dB_t^i + \sqrt{1-(\gamma^i)^2} d\tilde{B}_t^i) \end{aligned}$$

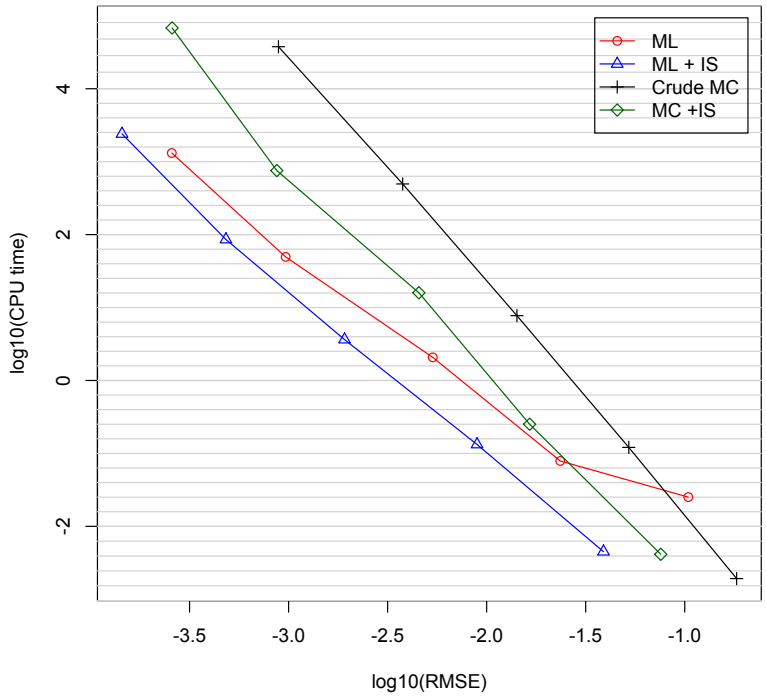


Figure 4.4.2: $\sqrt{MSE}$ vs. CPU time for a basket option in the local volatility model with $I = 5$, $r = 0.05$, $T = 1$, $S_0 = 100$, $K = 100$, $m = 4$.

where all the components of $B = (B^1, \dots, B^d)$ and $\tilde{B} = (\tilde{B}^1, \dots, \tilde{B}^d)$ are real valued Brownian motions. The vectors $\kappa = (\kappa^1, \dots, \kappa^d)$ and $a = (a^1, \dots, a^d)$ denote respectively the reversion rate and the mean level of each volatility process, while the vector ν is the volatility of the volatility process. The vector $\bar{\gamma} = (\gamma^1, \dots, \gamma^d)$ embodies the correlations between an asset and its volatility process, with $\gamma^i \in]-1, 1[$ for all $1 \leq i \leq d$. The vector valued processes B and $\tilde{B}$ are independent and satisfy

$$d\langle B \rangle_t = \Gamma_S \, dt \quad \text{and} \quad d\langle \tilde{B} \rangle_t = I_d \, dt$$

where we assume for our experiments that the covariance matrix Γ_S has the structure

$$\Gamma_S = \begin{pmatrix} 1 & \rho & \dots & \rho \\ \rho & 1 & \ddots & \vdots \\ \vdots & \ddots & \ddots & \rho \\ \rho & \dots & \rho & 1 \end{pmatrix} \tag{4.19}$$

with $\rho \in \left] \frac{-1}{I-1}, 1 \right[$, such that the matrix Γ_S is positive definite. The processes B and $\tilde{B}$ are Wiener processes with covariance matrices given by Γ_S and I_d respectively.

For the sake of simplicity, we decided not to add any extra correlation between the components of $\tilde{B}$, hence the choice $d\langle \tilde{B} \rangle = I_d \, dt$ and we assume in the following that all the γ^i's are equal, $\gamma^i = \gamma$ for $1 \leq i \leq d$. The correlations between the volatilities are entirely specified by the correlations between

the assets. Even though we do not aim at discussing the correlation structure of the multidimensional Heston model, we believe it is important to make precise the underlying correlation structure in the multidimensional model so that the experiments are easily reproducible.

The model can be equivalently written

$$dS_t^i = rS_t^i dt + \sqrt{\sigma_t^i} S_t^i dB_t^i$$
$$d\sigma_t^i = \kappa^i(a^i - \sigma_t^i)dt + \nu_t^i \sqrt{\sigma_t^i} dW_t^i$$

where the processes W and B are Wiener processes satisfying

$$d\langle B\rangle_t = \Gamma_S\, dt;\ d\langle B, W\rangle_t = \gamma\Gamma_S\, dt;\ d\langle W\rangle_t = (\gamma^2\Gamma_S + (1-\gamma^2)I_d)\, dt.$$

The process (B, W) with values in $\mathbb{R}^{2d}$ is a Wiener process with covariance matrix

$$\Gamma = \begin{pmatrix} \Gamma_S & \gamma\Gamma_S \\ \gamma\Gamma_S & \gamma^2\Gamma_S + (1-\gamma^2)I_d \end{pmatrix}.$$

Hence, the pair of processes (B, W) can be easily simulated by applying the Cholesky factorization of Γ to a standard Brownian motion with values in $\mathbb{R}^{2d}$.

Basket Option We consider a basket option as in the local volatility model. Figure 4.4.3 looks very much the same as in the case of the local volatility model (see Figure 4.4.2). The MLIS estimator always outperforms all the ML estimator by a factor of 3 to 4. Note that for small RMSE, the computational time can go beyond several hours, hence cutting it down by two or three times represents a real improvement.

Best of option We consider options with payoffs of the form $(\max_{1\le i\le d} S_T^i - K)_+$. The payoff of this option does obviously not satisfy the assumptions of Theorem 4.2.5 as the payoff of the "best of" options is not Hölder with $\alpha \ge 1$. Nonetheless, the multilevel approach beats the standard Monte Carlo technology by far (see Figure 4.4.4). Moreover, coupling importance sampling with the multilevel approach improves the accuracy. For a fixed RMSE, we can expect MLIS to be 3 faster that ML. This example shows the robustness of the method, which performs well whereas the theoretical assumptions are not satisfied.

4.5 Conclusion

We have presented a new estimator making the most of the recent works on multilevel Monte Carlo and on adaptive importance sampling. As expected, this new estimator outperforms the standard multilevel Monte Carlo estimator by a great deal. For a fixed accuracy measured in terms the mean squared error, the MLIS estimator is between 3 and 10 times faster that the standard multilevel Monte Carlo estimator. This efficiency of our MLIS approach could still be improved by monitoring the number of samples N'_ℓ to be used to approximate the variance v_{ℓ,N'_ℓ} in each level. Actually, we believe that there is no need to compute a too accurate approximation of this variance as a slight decrease in the accuracy of $\hat{\lambda}_\ell$ would not lead to a serious deterioration of the accuracy of the MLIS estimator but it could help to save a lot of computational time.

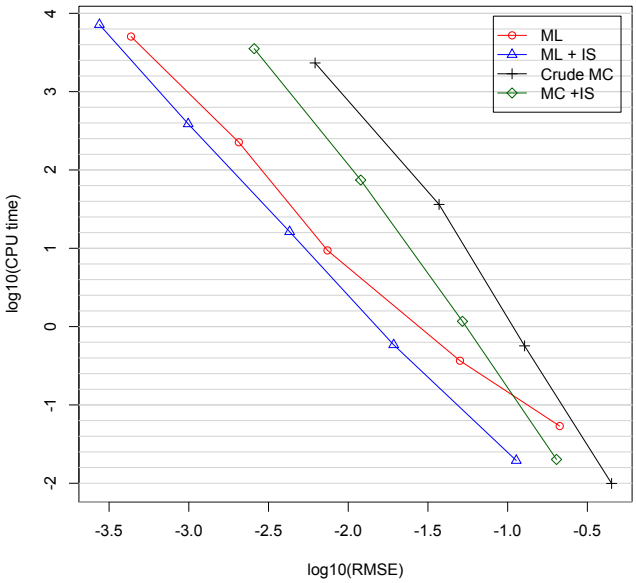


Figure 4.4.3: $\sqrt{MSE}$ vs. CPU time for a best of option in the multidimensional Heston model with $I = 10$, $r = 0.03$, $T = 1$, $S_0 = 100$, $K = 100$, $\nu = 0.01$, $\kappa = 2$, $a = 0.04$, $\gamma = -0.2$, $\rho = 0.3$ and $m = 4$.

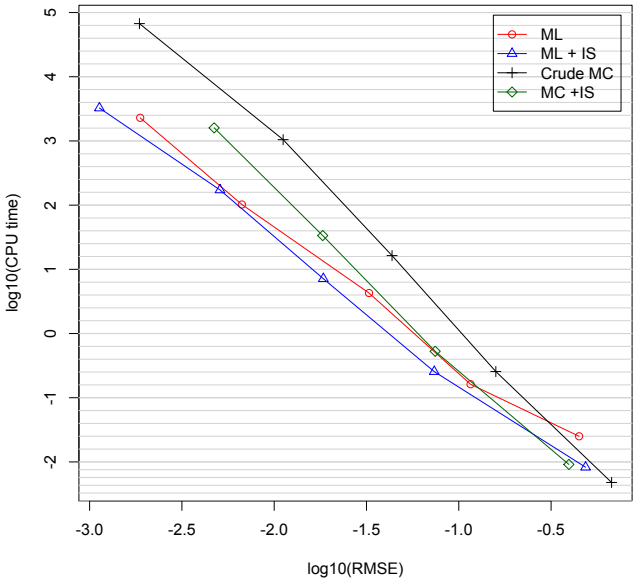


Figure 4.4.4: $\sqrt{MSE}$ vs. CPU time for a best of option in the multidimensional Heston model with $I = 5$, $r = 0.03$, $T = 1$, $S_0 = 100$, $K = 140$, $\nu = 0.25$, $\kappa = 2$, $a = 0.04$, $\gamma = 0.2$, $\rho = 0.5$ and $m = 4$.

Chapter 5

A stochastic optimization point of view to American options

This chapter summarizes my paper [L-21] on the use of stochastic optimisation to price American options using the dual representation of the price.

5.1 Introduction

The pricing of American options quickly becomes challenging as the dimension increases and the payoff gets complex. Many people have contributed to this problem usually by considering its dynamic programming principle formulation [88], [26], [89], [72], [23] and [5]. Among this so extensive literature, the practitioners seem to prefer the iterative optimal policy approach proposed by [72], which proves to be quite efficient in many situations. However, *true* path–dependent options cannot be handled by this approach. Solving the dynamic programming principle requires the computation of a conditional expectation, which is eventually handled by regression techniques. These techniques are known to suffer from the curse of dimensionality: global regression methods lead to high dimensional linear algebra problems, whereas the number of domains used by local methods blows up with the dimension. Despite the numerous parallel implementations of these techniques (see for instance [38]), we cannot expect to obtain a fully scalable algorithm. In this work, we follow the dual approach initiated by [80], and [31], which can naturally handle path dependent options. To make it implementable, we need a smart and finite dimensional approximation of the set of uniformly integrable martingales. We chose the set of truncated Wiener chaos expansions, which have some magic features in our problem: it regularizes the optimization problem and computing its conditional expectation exactly is straightforward. Then, the pricing problem boils down to a finite dimensional, convex and differentiable optimization problem. The optimization problem is solved using a *Sample Average Approximation* (see [83] and Chapter 3), which can be easily and efficiently implemented using parallel computing.

We fix some finite time horizon $T > 0$ and a filtered probability space $(\Omega, \mathcal{F}, (\mathcal{F}_t)_{0\leq t\leq T}, \mathbb{P})$, where $(\mathcal{F}_t)_{0\leq t\leq T}$ is supposed to be the natural augmented filtration of a $d-$dimensional Brownian motion B. On this space, we consider an adapted process $(S_t)_{0\leq t\leq T}$ with values in $\mathbb{R}^{d'}$ modeling a d'–dimensional underlying asset. The number of assets d' can be smaller than the dimension d of the Brownian motion to encompass the case of stochastic volatility models or stochastic interest rate. We assume that the short interest rate is modeled by an adapted process $(r_t)_{0\leq t\leq T}$ with values in $\mathbb{R}_+$ and that $\mathbb{P}$ is an associated risk neutral measure. We consider an adapted payoff process $\tilde{Z}$ and

introduce its discounted value process $\left(Z_t = \mathrm{e}^{-\int_0^t r_s ds}\,\tilde{Z}_t\right)_{0\leq t\leq T}$. We assume that the paths of Z are right continuous and that $\sup_{t\in[0,T]}|Z_t| \in L^2$. The process $\tilde{Z}$ can obviously take the simple form $(\phi(S_t))_{t\leq T}$ but it can also depend on the whole path of S up to the current time. So, our framework transparently deals with path–dependent option, which are far more difficult to handle using regression techniques.

We consider the American option paying $\tilde{Z}_t$ to its holder if exercised at time t. Standard arbitrage pricing theory defines the discounted time-t value of the American option to be

$$U_t = \operatorname{esssup}_{\tau\in\mathcal{T}_t} \mathbb{E}[Z_\tau|\mathcal{F}_{t_k}] \tag{5.1}$$

where $\mathcal{T}_t$ denotes the set of $\mathcal{F}-$stopping times with values in $[t,T]$. The integrability properties of Z ensure that U is a supermartingale of class (D) and hence has a Doob–Meyer decomposition

$$U_t = U_0 + M_t^\star - A_t^\star \tag{5.2}$$

where $M^\star$ is a martingale vanishing at zero and $A^\star$ is a predictable integrable increasing process also vanishing at zero. With our assumptions on Z, $M^\star$ is square integrable. [48] and [80] found an alternative representation of the price at time-0 of the American option as the minimum value of the following optimization problem

$$U_0 = \inf_{M\in H_0^2} \mathbb{E}\left[\sup_{t\leq T}(Z_t - M_t)\right] = \mathbb{E}\left[\sup_{t\leq T}(Z_t - M_t^\star)\right] \tag{5.3}$$

where H_0^2 denotes the set of square integrable martingales vanishing at zero. A martingale reaching the infimum is called an *optimal* martingale. As the dual price problem writes as a convex minimisation problem, the set of all optimal martingales is a convex subset of H_0^2. Among the martingales reaching the infimum in (5.3), some of them actually satisfy the pathwise equality $\sup_{t\leq T} Z_t - M_t = U_0$. These martingales are called *surely optimal*. Any surely optimal martingale reaches the lower bound in (5.3) but not all optimal martingales are surely optimal. We refer to [85] for a detailed characterisation of optimal martingales. Anyway, [54] proved the uniqueness of surely optimal martingales within the continuing region, ie. for any surely optimal martingale M and any optimal strategy τ, $(M_{t\wedge\tau})_t = (M^\star_{t\wedge\tau})_t$ a.s.

The most famous method using the dual representation (5.3) is probably the primal–dual approach of [2], which heavily relies on the knowledge of an optimal exercising policy. The a priori knowledge may take the form of nested Monte Carlo simulations as in [84], and [63]. To circumvent this difficulty, [81] explained how to construct a *good* martingale. In a Wiener framework, [11] investigated this approach by relying on the martingale representation theorem to build *good* martingales. When trying to practically use the dual formulation (5.3), the first difficulty is to find a rich enough but finite dimensional approximation of H_0^2 and then we face a finite although potentially high–dimensional minimization problem (see [10] for one way of handling this approach).

The minimization problem (5.3) can be equivalently formulated as

$$U_0 = \inf_{X\in L_0^2(\Omega,\mathcal{F}_T,\mathbb{P})} \mathbb{E}\left[\sup_{0\leq t\leq T}(Z_t - \mathbb{E}[X|\mathcal{F}_t])\right] \tag{5.4}$$

where $L_0^2(\Omega,\mathcal{F}_T,\mathbb{P})$ is the set of square integrable $\mathcal{F}_T-$ random variables with zero mean. In this work, we suggest to use the truncated Wiener chaos expansion as a finite dimensional approximation of $L^2(\Omega,\mathcal{F}_T,\mathbb{P})$. Since Wiener chaos are orthogonal for the L^2 inner product, the computations of

the conditional expectations $\mathbb{E}[X|\mathcal{F}_t]$ become straightforward and boil down to dropping some terms in the chaos expansion, which makes our approach very convenient. Based on this approximation, we propose a scalable algorithm and study its convergence.

The chapter starts with a presentation of the Wiener chaos expansion and some of its useful properties in Section 5.2. Then, we can develop the core of our work in Section 5.3, in which we explain how the price of the American option can be approximated by the solution of a finite dimensional optimization problem. First, we analyze the properties of the optimization problem in order to prove the convergence of its solution to the American option price. Second, we study its sample average approximation, which makes the problem tractable, and prove its convergence. Based on all these theoretical results, we present our algorithm in Section 5.4 and discuss its parallel implementation on distributed memory architectures. Finally, some numerical examples are presented in Section 5.5.

Notation

- For $\alpha \in \mathbb{N}^q$, $|\alpha|_1 = \sum_{i=1}^q \alpha_i$.
- For $n \geq 1$, $0 = t_0 < t_1 < \cdots < t_n = T$ is a time grid of $[0,T]$ satisfying $\lim_{n\to\infty} \sup_{0\leq k\leq n-1} |t_{k+1} - t_k| = 0$.
- For $n \geq 1$, the discrete time filtration $\mathcal{G}$ is defined by $\mathcal{G}_k = \sigma(B_{t_{i+1}} - B_{t_i}, i = 0, \ldots, k-1)$ for all $1 \leq k \leq n$, while $\mathcal{G}_0$ is the trivial sigma algebra. Obviously, $\mathcal{G}_k \subset \mathcal{F}_{t_k}$ for all $0 \leq k \leq n$.
- For $1 \leq q \leq d$, $\mathbb{I}(r) \in \{0,1\}^n$ denotes the vector $(\underbrace{0,\ldots,0}_{r-1}, 1, \underbrace{0,\ldots,0}_{n-r})$.
- For $1 \leq q \leq d$, and $1 \leq r \leq n$, $\mathbb{I}(r,q) \in \mathbb{N}^{n\times d}$ with all components equal to 0 except the component with index (r,q) which is equal to 1.

We recall some useful definitions related to Malliavin calculus using the notation of [75].

- Let $\mathcal{S}$ denote the class of smooth random variables of the form $F = f(W(h_1), \cdots, W(h_\ell))$ where $m \geq 1$, $f \in C_p^\infty(\mathbb{R}^{\ell\times d}, \mathbb{R})$, for all $j \leq \ell$, $h_j = (h_j^1, \cdots, h_j^d) \in L^2([0,T], \mathbb{R}^d)$ and for all $i \leq d$, $W^i(h_j^i) = \int_0^T h_j^i(t) dW_t^i$.
- For $F \in \mathcal{S}$, the Malliavin derivative of F denoted by $DF = (D^1, \cdots D^d)$ is a stochastic process with values in $\mathbb{R}^d$. For $t \leq T$ and $1 \leq i \leq d$, D_t^i is defined by

$$D_t^i F = \sum_{j=1}^{\ell} \partial_j f(W(h_1), \cdots, W(h_m)) h_j^i(t).$$

  With this notation, D_t is a gradient operator.

- For $m \geq 1$, a multi–index $\alpha \in \{1, \cdots, d\}^m$ and a tuple of dates $(t_1, \cdots, t_m)$, we write

$$D^\alpha_{t_1,\cdots,t_m} F = D^{\alpha_1}_{t_1}(\cdots(D^{\alpha_m}_{t_m} F)).$$

  $D^{(m)}F = \{D^\alpha_{t_1,\cdots,t_m} F \ : \ \alpha \in \{1,\cdots,d\}^m, \ (t_1,\cdots,t_m) \in [0,T]^m\}$ can be seen as a measurable function defined on $\Omega \times [0,T]^m$. When $\alpha_1 = \cdots = \alpha_m = 1$, we drop the multi–index and just write $D_{t_1,\cdots,t_m}$.

- Let $\mathbb{D}^{m,2}$ be the closure of $\mathcal{S}$ w.r.t. the following norm

$$\|F\|^2_{\mathbb{D}^{m,2}} = \mathbb{E}[|F|^2] + \sum_{r=1}^{m} \sum_{|\alpha|_1=r} \mathbb{E}\left[\int_{[0,T]^r} \left|D^{\alpha}_{t_1,\cdots,t_r} F\right|^2 dt_1 \cdots dt_r\right].$$

5.2 Wiener chaos expansion

In this section, we recall some well known material about Wiener chaos expansion using the Hermite polynomial point of view and state some results on the Malliavin derivative of a chaos expansion. We refer the reader to [75] for further details.

5.2.1 The one–dimensional framework

For the sake of clearness, we first present the Wiener chaos expansion in the case $d = 1$ (ie. B is a real valued Brownian motion).

Let H_i be the $i-th$ Hermite polynomial defined by

$$H_0(x) = 1; \qquad H_i(x) = (-1)^i \, \mathrm{e}^{x^2/2} \, \frac{d^i}{dx^i}(\mathrm{e}^{-x^2/2}), \text{ for } i \geq 1. \tag{5.5}$$

They satisfy for all integer i, $H_i' = H_{i-1}$ with the convention $H_{-1} = 0$. We recall that if (X, Y) is a random normal vector with $\mathbb{E}[X] = \mathbb{E}[Y] = 0$ and $\mathbb{E}[X^2] = \mathbb{E}[Y^2] = 1$

$$\mathbb{E}[H_i(X)H_j(Y)] = i! \, (\mathbb{E}[XY])^i \; \mathbf{1}_{i=j}. \tag{5.6}$$

For all $p \geq 0$, we define the spaces

$$\mathcal{H}_p = \mathrm{span}\left\{H_p\left(\int_0^T f_t dB_t\right) \; : \; f \in L^2([0,T])\right\}. \tag{5.7}$$

It is well known that $L^2(\Omega, \mathcal{F}_T, \mathbb{P}) = \bigoplus_{p=0}^{\infty} \mathcal{H}_p$, see [75, Theorem 1.1.1]. Let us introduce the generalized Hermite polynomials defined for any multi–index $\alpha = (\alpha_i)_{i\geq 1} \in \mathbb{N}^{\mathbb{N}}$

$$\widehat{H}_{\alpha}(x) = \prod_{i\geq 1} H_{\alpha_i}(x_i), \quad \text{for } x \in \mathbb{R}^{\mathbb{N}}. \tag{5.8}$$

If $(f_i)_{i\geq 0}$ is an orthonormal basis of $L^2([0,T])$, then the random variables

$$\left\{\widehat{H}_{\alpha}\left(\left(\int_0^T f_i(t) dB_t\right)_{i\geq 0}\right) \; : \; \alpha \in \mathbb{N}^{\mathbb{N}}, \; |\alpha|_1 \leq p\right\}$$

form a complete orthonormal system in $\mathcal{H}_p$, see [75, Proposition 1.1.1].

Consider the indicator functions of the grid $0 = t_0 < t_1 < \ldots. < t_n = T$ defined by

$$f_i(t) = \mathbf{1}_{]t_{i-1},t_i]}(t)/\sqrt{t_i - t_{i-1}}, \quad i = 1, \ldots, n, \tag{5.9}$$

With this choice for the $(f_i)_i$,

$$\int_0^T f_i(t) dB_t = \frac{B_{t_i} - B_{t_{i-1}}}{\sqrt{t_i - t_{i-1}}} = G_i.$$

Note that the random variables G_i are i.i.d. following the standard normal distribution. We denote by $\mathcal{C}_{p,n}$ the vector space generated by the random variables

$$\left\{\widehat{H}_\alpha\left(G_1,\cdots,G_n\right)\ :\ \alpha\in A_{p,n}\right\}$$

where $A_{p,n}=\{\alpha\in\mathbb{N}^n\ :\ |\alpha|_1\leq p\}$. For a random variable $F\in L^2(\Omega,\mathcal{F}_T,\mathbb{P})$ we define its truncated chaos expansion of order p as its projection on $\mathcal{C}_{p,n}$ and write

$$C_{p,n}(F)=\sum_{\alpha\in A_{p,n}}\lambda_\alpha\widehat{H}_\alpha(G_1,\ldots,G_n)$$

5.2.2 The multi–dimensional framework

Now, we are back to our original multi–dimensional setting, as explained in Section 5.1. The process B is a Brownian motion with values in $\mathbb{R}^d$. The natural way to extend the Hermite polynomial expansion to a higher dimensional setting is to consider a tensor product of Hermite polynomials evaluated on a tensor basis of $L^2([0,T],\mathbb{R}^d)$.

Consider the functions $(f_i)_i$ with values in $\mathbb{R}^d$ defined by

$$f_i^j(t)=\frac{\mathbf{1}_{]t_{i-1},t_i]}(t)}{\sqrt{t_i-t_{i-1}}}\mathbf{e}_j,\ i=1,\ldots,n,\ j=1,\ldots,d$$

where $(\mathbf{e}_1,\ldots,\mathbf{e}_d)$ denotes the canonical basis of $\mathbb{R}^d$. The $p-th$ Wiener chaos $\mathcal{C}_{p,n}$ is defined as the vector space generated by the random variables

$$\left\{\prod_{j=1}^d\widehat{H}_{\alpha^j}(G_1^j,\ldots,G_n^j)\ :\ \alpha\in A_{p,n}^{\otimes d}\right\}$$

where $G_i^j=\frac{B_{t_i}^j-B_{t_{i-1}}^j}{\sqrt{t_i-t_{i-1}}}$ and $A_{p,n}^{\otimes d}=\left\{\alpha\in(\mathbb{N}^n)^d\ :|\alpha|_1\leq p\right\}$. Using the independence of the Brownian increments and the orthogonality of the Hermite polynomials, the truncated chaos expansion of a square integrable random variable F is given by

$$C_{p,n}(F)=\sum_{\alpha\in A_{p,n}^{\otimes d}}\lambda_\alpha\widehat{H}_\alpha^{\otimes d}(G_1,\ldots,G_n)$$

where $\widehat{H}_\alpha^{\otimes d}(G_1,\ldots,G_n)=\prod_{j=1}^d\widehat{H}_{\alpha^j}(G_1^j,\ldots,G_n^j)$, $\forall\alpha\in(\mathbb{N}^n)^d$. With an obvious abuse of notation, we write, for $\lambda\in\mathbb{R}^{A_{p,n}^{\otimes d}}$,

$$C_{p,n}(\lambda)=\sum_{\alpha\in A_{p,n}^{\otimes d}}\lambda_\alpha\widehat{H}_\alpha^{\otimes d}(G_1,\ldots,G_n).$$

We also introduce the set of multi–indices truncated after time t_k

$$A_{p,n}^{\otimes d,k}=\left\{\alpha\in A_{p,n}^{\otimes d}\ :\ \forall j\in\{1,\ldots,d\},\ \forall\ell>k,\ \alpha_\ell^j=0\right\}. \tag{5.10}$$

Proposition 5.2.1 *Let F be a real valued random variable in $L^2(\Omega,\mathcal{F}_T,\mathbb{P})$ and let $k\in\{1,\ldots,n\}$ and $p\geq 0$*

$$\mathbb{E}[C_{p,n}(F)|\mathcal{F}_{t_k}]=\sum_{\alpha\in A_{p,n}^{\otimes d,k}}\lambda_\alpha\ \widehat{H}_\alpha^{\otimes d}(G_1,\ldots,G_n).$$

Remark 5.2.2 Since the sum appearing in $\mathbb{E}[C_{p,n}(F)|\mathcal{F}_{t_k}]$ is reduced to a sum over the set of multi–indices $\alpha \in A^k_{p,n}$, it actually only depends on the first k increments $(G_1, \ldots, G_k)$. One can easily check that $\mathbb{E}[C_{p,n}(F)|\mathcal{F}_{t_k}]$ is actually given by the chaos expansion of F on the first k Brownian increments. Hence, computing a conditional expectation simply boils down to dropping terms. While it may look like a naive way to proceed, it is indeed correct in our setting.

Remark 5.2.3 The discrete time sequence $(\mathbb{E}[C_{p,n}(F)|\mathcal{F}_{t_k}])_{0\leq k\leq n}$ is of course adapted to the filtration $(\mathcal{F}_{t_k})_k$ but also to the smaller filtration $(\mathcal{G}_k)_k$. This property plays a crucial when approximating a random variable $F \in L^2(\Omega, \mathcal{G}_n, \mathbb{P})$ as we know that in such a finite dimensional setting $\lim_{p\to\infty} C_{p,n}(F) = F$ in the L^2-sense. This result holds for the fixed value n. If F were only $\mathcal{F}_T-$measurable and not $\mathcal{G}_n-$measurable, we would need to impose that $F \in \mathbb{D}^{1,2}$ to obtain $\lim_{p\to\infty,n\to\infty} C_{p,n}(F) = F$. In this latter case, it is required to let n go to infinity to recover F.

Proposition 5.2.4 *Let F be a real valued random variable in $L^2(\Omega, \mathcal{F}_T, \mathbb{P})$ and let $k \in \{1, \ldots, n\}$ and $p \geq 1$. For $t > t_k$, $D_t\mathbb{E}[C_{p,n}(F)|\mathcal{F}_{t_k}] = 0$.*

For all $t \in]t_{r-1}, t_r]$ with $1 \leq r \leq k$, and $q = 1, \ldots, d$,

$$D^q_t\mathbb{E}[C_{p,n}(F)|\mathcal{F}_{t_k}] = \frac{1}{\sqrt{t_r - t_{r-1}}} \sum_{\alpha\in A^{\otimes d,k}_{p,n}, \alpha^q_r\geq 1} \lambda_\alpha \, \widehat{H}^{\otimes d}_{\alpha-\mathbb{I}(r,q)}(G_1, \ldots, G_n)$$

where $(\alpha - \mathbb{I}(r,q))^j_i = \alpha^j_i - \mathbf{1}_{j=q,i=r}$.

Remark 5.2.5 The Malliavin derivative of a chaos expansion still writes as a chaos expansion and hence is a Hermite polynomial of Brownian increments. The roots of a non zero polynomial being a zero measure set and since the Brownian increments have a joined density, the Malliavin derivative of a chaos expansion is almost surely non zero as soon as one of the coefficients λ_α is non zero for $\alpha \in A^{\otimes d,k}_{p,n}$ such that $\alpha^j_r \geq 1$ for some $j \in \{1, \ldots, d\}$.

For $i, k \in \{1, \ldots, n\}$, with $i < k$, we introduce the set $A^{\otimes d,i:k}_{p,n}$ defined as $A^{\otimes d,k}_{p,n} \setminus A^{\otimes d,i}_{p,n}$.

$$A^{\otimes d,i:k}_{p,n} = \Big\{\alpha \in (\mathbb{N}^n)^d \ : \ |\alpha|_1 \leq p, \text{ and } \forall 1 \leq j \leq d, \ \forall \ell \notin \{i+1, \ldots, k\}, \ \alpha^j_\ell = 0\Big\}. \tag{5.11}$$

5.3 Pricing American options using Wiener chaos expansion and sample average approximation

In this section, we aim at approximating the dual price (5.4) by a tractable optimization problem. This involves two kinds of approximations: first, to approximate the space $L^2_0(\Omega, \mathcal{F}_T, \mathbb{P})$ by a finite dimensional vector space; second, to replace the expectation by a sample average approximation.

The dual price writes

$$\inf_{X\in L^2_0(\Omega,\mathcal{F}_T,\mathbb{P})} \mathbb{E}\left[\sup_{0\leq t\leq T}(Z_t - \mathbb{E}[X|\mathcal{F}_t])\right].$$

In this optimization problem, we replace X by its chaos expansion $C_{p,n}(X)$, which has no constant term as $\mathbb{E}[X] = 0$ and we approximate the supremum by a discrete time maximum. Then, we face a finite dimensional minimization problem to determine the optimal solution within the subset $\mathcal{C}_{p,n}$

$$\inf_{\lambda\in\mathbb{R}^{A^{\otimes d}_{p,n}}, \, \lambda_0=0} \mathbb{E}\left[\max_{0\leq k\leq n}(Z_{t_k} - \mathbb{E}[C_{p,n}(\lambda)|\mathcal{F}_{t_k}])\right]. \tag{5.12}$$

In Section 5.3.1, we prove that this optimization problem is convex, has a solution (see Proposition 5.3.1) and converges to the price of the American option (see Proposition 5.3.2). Moreover, as the cost function is differentiable, any minimizer is a zero of the gradient (see Proposition 5.3.4).

To come up with a fully implementable algorithm, Section 5.3.2 presents the sample average approximation of (5.12), which consists in replacing the expectation by a Monte Carlo summation. We prove in Proposition 5.3.5 that the solution of the sample average approximation converges to the solution of (5.12) when the number of samples goes to infinity.

5.3.1 A stochastic optimization approach

We fix $p \geq 1$ and define the random functions $v_{p,n}(\cdot,\cdot;Z,G) : \mathbb{R}^{A_{p,n}^{\otimes d}} \times \{0,\dots,n\}$ by

$$v_{p,n}(\lambda,k;Z,G) = Z_{t_k} - \sum_{\alpha\in A_{p,n}^{\otimes d}} \lambda_\alpha \mathbb{E}\left[\widehat{H}_\alpha^{\otimes d}(G_1,\dots,G_n)\,\middle|\,\mathcal{F}_{t_k}\right],$$

With the help of Proposition 5.2.1, the random functions $v_{p,n}$ can be written

$$v_{p,n}(\lambda,k,Z,G) = Z_{t_k} - \sum_{\alpha\in A_{p,n}^{\otimes d,k}} \lambda_\alpha \widehat{H}_\alpha^{\otimes d}(G_1,\dots,G_n). \tag{5.13}$$

We consider the cost function $V_{p,n} : \mathbb{R}^{A_{p,n}^{\otimes d}} \to \mathbb{R}$ defined by

$$V_{p,n}(\lambda) = \mathbb{E}\left[\max_{0\leq k\leq n} v_{p,n}(\lambda,k;Z,G)\right] \tag{5.14}$$

and we approximate the solution of (5.4) by

$$\inf_{\lambda\in\mathbb{R}^{A_{p,n}^{\otimes d}},\ \lambda_0=0} V_{p,n}(\lambda). \tag{5.15}$$

5.3.1.1 Convergence results

Proposition 5.3.1 *The minimization problem* (5.15) *has at least one solution.*

Proof. As the supremum of linear functions is convex, the random function $\lambda \longmapsto \max_{k\leq n} v_{p,n}(\lambda,t_k,Z,G)$ is almost surely convex. The convexity of $V_{p,n}$ ensues from the linearity of the expectation.

Let us prove that $V_{p,n}(\lambda) \to \infty$ when $|\lambda| \to \infty$. Note that $V_{p,n}(\lambda) \geq \mathbb{E}\left[(C_{p,n}(\lambda))_-\right] \geq \frac{1}{2}\mathbb{E}\left[|C_{p,n}(\lambda)|\right]$, where we have used that $|x| = 2x_- + x$ and $\mathbb{E}[C_{p,n}(\lambda)] = 0$.

$$\mathbb{E}\left[|C_{p,n}(\lambda)|\right] = |\lambda|\,\mathbb{E}\left[|C_{p,n}(\lambda/|\lambda|)|\right] \geq |\lambda| \inf_{\mu\in\mathbb{R}^{A_{p,n}^{\otimes d}},|\mu|=1} \mathbb{E}\left[|C_{p,n}(\mu)|\right]. \tag{5.16}$$

By a standard continuity argument, the infimum is attained. Moreover, it is strictly positive as otherwise there would exist $\mu \in \mathbb{R}^{A_{p,n}^{\otimes d}}$ with $|\mu| = 1$ s.t. $\mathbb{E}\left[|C_{p,n}(\mu)|\right] = 0$. Using the orthogonality of the family $\left(H_\alpha^{\otimes d}\right)_{\alpha\in A_{p,n}^{\otimes d}}$, we would immediately deduce that $\mu = 0$. Hence, we show that $V_{p,n}(\lambda) \to \infty$ when $|\lambda| \to \infty$. The growth at infinity of $V_{p,n}$ combined with its convexity yields the existence of a solution to the minimization problem (5.15). ■

Proposition 5.3.1 ensures the existence of $\lambda^\sharp_{p,n}$ solving (5.15), ie.

$$V_{p,n}(\lambda^\sharp_{p,n}) = \inf_{\lambda \ s.t. \ \lambda_0=0} V_{p,n}(\lambda). \tag{5.17}$$

To study the convergence of the $V_{p,n}(\lambda^\sharp_{p,n})$, we introduce the Bermudan option with exercising dates $t_0, \dots, t_n$ and with discounted payoff $(Z_{t_k})_k$. Let $\hat{U}^n_k$ be its time-t_k price. The sequence $(\hat{U}^n_k)_{0\le k\le n}$ is a supermartingale admitting the Doob–Meyer decomposition $\hat{U}^n_k = \hat{U}^n_0 + \hat{M}^{\star,n}_k - \hat{A}^{\star,n}_k$ where $\hat{M}^n$ is a square integrable $(\mathcal{F}_{t_k})_k-$martingale and $\hat{A}^n$ a predictable increasing process for the filtration $(\mathcal{F}_{t_k})_k$. The time-0 price can be expressed as

$$\hat{U}^n_0 = \inf_{X\in L^2_0(\Omega,\mathcal{F}_T,\mathbb{P})} \mathbb{E}\left[\max_{0\le k\le n}(Z_{t_k} - \mathbb{E}[X|\mathcal{F}_{t_k}])\right] = \mathbb{E}\left[\max_{0\le k\le n}(Z_{t_k} - M^{\star,n}_{t_k})\right]. \tag{5.18}$$

Note that $V_{p,n}(\lambda^\sharp_{p,n}) \ge \hat{U}^n_0$.

Proposition 5.3.2 *We have*

$$\left|V_{p,n}(\lambda^\sharp_{p,n}) - U_0\right| \le 2\left\|M^\star_T - C_{p,n}(M^\star_T)\right\|_2 + \mathbb{E}\left[\max_k(Z_{t_k} - M^\star_{t_k})\right] - \mathbb{E}\left[\sup_t(Z_t - M^\star_t)\right].$$

Moreover, assume that $\hat{U}^n_0$ converges to U_0 with n. Then, $V_{p,n}(\lambda^\sharp_{p,n})$, converges to U_0 when both p and n go to infinity.

Note that $\mathbb{E}\left[\max_k(Z_{t_k} - M^\star_{t_k})\right] \ge \hat{U}^n_0$, hence

$$\mathbb{E}\left[\max_k(Z_{t_k} - M^\star_{t_k})\right] - \mathbb{E}\left[\sup_t(Z_t - M^\star_t)\right] \le U_0 - \hat{U}^n_0.$$

We refer to [27, 67] for results on the convergence of $\hat{U}^n_0$ to U_0. The convergence of $\left\|M^{\star,n}_T - C_{p,n}(M^\star_T)\right\|_2$ to 0 when p, n go to infinity ensues from [75, Theorem 1.1.1, Proposition 1.1.1].

Corollary 5.3.3 *Assume the discounted payoff $(Z_{t_k})_k$ of the Bermudan option is $\mathcal{G}-$adapted. Then, $V_{p,n}(\lambda^\sharp_{p,n})$ converges to the price of the Bermudan option when p goes to infinity.*

5.3.1.2 Regularity of the optimization problem

Most convex optimization algorithms mainly rely on the gradient of the cost function. We end this section by proving that $V_{p,n}$ is almost everywhere differentiable, which implies that $\nabla V_{p,n}(\lambda^\sharp_{p,n}) = 0$. We introduce the set of random indices for which the pathwise maximum is attained

$$\mathcal{I}(\lambda, Z, G) = \left\{0 \le k \le n \ : \ v_{p,n}(\lambda, k; Z, G) = \max_{\ell\le n} v_{p,n}(\lambda, \ell; Z, G)\right\}.$$

Proposition 5.3.4 *Let $p \ge 1$. Assume that*

$$\forall 1 \le r \le k \le n, \ \forall F \ \mathcal{F}_{t_k} - \textit{measurable}, \ F \in \mathcal{C}_{p-1,n}, \ F \ne 0, \ \exists \ q' \in \{1, \dots, d\} \ \textit{s.t.}$$
$$\mathbb{P}\left(\forall t \in]t_{r-1}, t_r], \ D^{q'}_t Z_{t_k} + F = 0 \mid Z_{t_k} > 0\right) = 0. \tag{5.19}$$

Define the open set

$$\Lambda = \{(\lambda_\alpha)_\alpha \in \mathbb{R}^{A_{p,n}^{\otimes d}} \ : \ \forall r \in \{1,\dots,n\},\ \exists(\alpha, q \in \{1,\dots,d\})\ \textit{s.t.}\ \alpha_r^q \geq 1\ \textit{and}\ \lambda_\alpha \neq 0\}.$$

Then, the function $V_{p,n}$ is differentiable on the set Λ and the gradient $\nabla V_{p,n}$ is given by

$$\nabla V_{p,n}(\lambda) = \mathbb{E}\left[\mathbb{E}\left[\widehat{H}^{\otimes d}(G_1,\dots,G_n) \mid \mathcal{F}_{t_i}\right]_{|\{i\}=\mathcal{I}(\lambda,Z,G)}\right].$$

Sketch of the proof. For all Z and G, the function $\lambda \longmapsto \max_{k\leq n} v_{p,n}(\lambda,k,Z,G)$ is subdifferentiable. It ensues from [17] that the subdifferential $\partial V_{p,n}(\lambda)$ writes as the expectation of the subdifferential of its integrand

$$\partial V_{p,n}(\lambda) = \left\{ \mathbb{E}\left[\sum_{i\in\mathcal{I}(\lambda,Z,G)} \beta_i \mathbb{E}[\widehat{H}^{\otimes d}(G_1,\dots,G_n)|\mathcal{F}_{t_i}]\right] \ : \ \beta_i \geq 0,\ \mathcal{F}_T - meas.,\ \sum_i \beta_i = 1 \right\}.$$

It is sufficient to prove for any λ with no zero component, the set $\mathcal{I}(\lambda,Z,G)$ is almost surely reduced to a single value as in this case the subdifferential $\partial V_{p,n}(\lambda)$ contains a unique element, which is then the gradient.

From (5.19) and [75, Theorem 2.1.3], we can prove that for any $r < k$, and any $F \in \mathcal{C}_{p-1,n}$, $Z_{t_k} - Z_{t_r} + F$ has a density, which yields the differentiability of $V_{p,n}$. ■

5.3.2 The Sample Average Approximation point of view

From Proposition 5.3.2, we can approximate U_0 by solving the minimization problem (5.15), which admits at least one solution $\lambda_{p,n}^\sharp$, ie.

$$V_{p,n}(\lambda_{p,n}^\sharp) = \inf_{\lambda\in A_{p,n}^{\otimes d},\ \lambda_0=0} V_{p,n}(\lambda)$$

where $V_{p,n}$ defined by (5.14) is an expectation, which is barely tractable. To practically solve such a problem, two different approaches are commonly used. Either, one uses a stochastic algorithm or one replaces the expectation by a sample average approximation. In this work, we target large problems, which puts scalability as a primary requirement. The intrinsic sequential nature of stochastic algorithms has led us to prefer the sample average approximation approach. Moreover, we are more interested in the value function at the minimum rather than in its minimizer and unlike stochastic algorithm, standard optimization algorithms provide both at once.

We introduce the sample average approximation of $V_{p,n}$ defined by

$$V_{p,n}^m(\lambda) = \frac{1}{m}\sum_{i=1}^m \max_{0\leq k\leq n} v_{p,n}(\lambda,k;Z^{(i)},G^{(i)})$$

where $(Z^{(i)},G^{(i)})_{1\leq i\leq m}$ are i.i.d samples from the distribution of (Z,G).

For large enough m, $V_{p,n}^m$ inherits from the smoothness of $V_{p,n}$ and is in particular convex and a.s. differentiable at any point with no zero component. Then, we easily deduce from Proposition 5.3.1 that there exits $\lambda_{p,n}^m$ such that

$$V_{p,n}^m(\lambda_{p,n}^m) = \inf_{\lambda\in\mathbb{R}^{A_{p,n}^{\otimes d}},\ \lambda_0=0} V_{p,n}^m(\lambda)$$

and moreover $\nabla V_{p,n}^m(\lambda_{p,n}^m) = 0$. The main difficulty in studying the convergence of $V_{p,n}^m(\lambda_{p,n}^m)$ when m goes to infinity comes from the non compactness of the set $\mathbb{R}^{A_{p,n}^{\otimes d}}$. To circumvent this problem, we adapt to non strictly convex problems the technique used in [L-12].

Proposition 5.3.5 *The sequence $V_{p,n}^m(\lambda_{p,n}^m)$ converges a.s. to $V_{p,n}(\lambda_{p,n}^\sharp)$ when $m \to \infty$. Moreover, the distance between $\lambda_{p,n}^m$ and the convex set of minimizers in* (5.15) *converges to zero as m goes to infinity.*

The proof of Proposition 5.3.5 very much looks like the proof of Proposition 3.2.5 with the strong convexity assumption replaced by the coercivity condition (5.16).

Although $V_{p,n}^m$ is not twice differentiable and the classical central limit theorem for sample average approximations cannot be applied, we can study the variance of $V_{p,n}^m(\lambda_{p,n}^m)$ and we obtain some asymptotic bounds. Before stating our result, we introduce, for $\lambda \in \mathbb{R}^{A_{p,n}^{\otimes d}}$, the notation $M_k(\lambda) = \mathbb{E}[C_{p,n}(\lambda)|\mathcal{F}_{t_k}]$ for $0 \le k \le n$. We write $M_k^{(i)}(\lambda)$ for the value computed using the sample $G^{(i)}$.

Proposition 5.3.6 *Assume $\lambda_{p,n}^\sharp$ is unique. Then,*

$$\frac{1}{m}\sum_{i=1}^{m}\left(\max_{0\le k\le n} Z_{t_k}^{(i)} - M_k^{(i)}(\lambda_{p,n}^m)\right)^2 - V_{p,n}^m(\lambda_{p,n}^m)^2$$

is a convergent estimator of $\mathrm{Var}(\max_{k\le 0\le n} Z_{t_k} - M_k(\lambda_{p,n}^\sharp))$ *and moreover if $\lambda_{p,n}^m$ is bounded,* $\lim_{m\to\infty} m\,\mathrm{Var}\left(V_{p,n}^m(\lambda_{p,n}^m)\right) = \mathrm{Var}(\max_{k\le 0\le n} Z_{t_k} - M_k(\lambda_{p,n}^\sharp))$.

Proposition 5.3.6 enables us to monitor the variance of our estimator online as for a standard Monte Carlo estimator. Even though the terms involved in $V_{p,n}^m(\lambda_{p,n}^m)$ are not independent, the classical variance estimator gives the right result. In practice, one should not feel concerned with the boundedness condition used in the proposition as we know from the proof of Proposition 5.3.5 that for large enough m we can impose a compactness constraint to the optimization problem without changing its result. Hence, one can pragmatically rely on the proposed variance estimator.

5.4 The algorithm

Any optimization algorithm requires to repeatedly compute $V_{p,n}^m$ and therefore the truncated chaos expansion, which becomes the most time consuming part of our approach as the dimension and/or p increase. A lot of computational time can be saved by considering slightly modified martingales, which only start the first time the option goes in the money.

5.4.1 An improved set of martingales

We define the first time the option goes in the money by

$$\tau_0 = \inf\{k \ge 0 \ : \ Z_{t_k} > 0\} \wedge n,$$

which is a $\mathcal{F}-$ stopping time and becomes a $\mathcal{G}-$ stopping time when the sequence $(Z_{t_k})_k$ is $\mathcal{G}-$ adapted. To consider martingales only starting once the option has been in the money, we define

$$N_k(\lambda) = \sum_{\ell=1}^{k}(M_\ell(\lambda) - M_{\ell-1}(\lambda))\mathbf{1}_{\ell-1\ge\tau_0} = (M_k(\lambda) - M_{\tau_0}(\lambda))\mathbf{1}_{k>\tau_0} = M_k(\lambda) - M_{k\wedge\tau_0}(\lambda).$$

We easily check that $N(\lambda)$ is a $(\mathcal{F}_{t_k})_{0\le k\le n}-$ martingale. It is clear from the proof proposed by [80] that in the dual price of a Bermudan option (see (5.3)) the maximum can be shrunk to the random interval $[\tau_0, n]$. Hence, it is sufficient to consider

$$\inf_{\lambda\in\mathbb{R}^{A_{p,n}^{\otimes d}},\, \lambda_0=0} \mathbb{E}\left[\max_{\tau_0\le k\le n}(Z_{t_k} - M_k(\lambda))\right].$$

Using Doob's stopping theorem, we have, for any fixed λ,

$$\mathbb{E}\left[\max_{\tau_0\le k\le n}(Z_{t_k} - M_k(\lambda))\right] = \mathbb{E}\left[\max_{\tau_0\le k\le n}(Z_{t_k} - (M_k(\lambda) - M_{\tau_0}(\lambda)))\right] = \mathbb{E}\left[\max_{\tau_0\le k\le n}(Z_{t_k} - N_k(\lambda))\right].$$

We deduce from this equality that minimizing over either set of martingales $M(\lambda)$ or $N(\lambda)$ leads to the same minimum value and that both problems share the same properties, which justifies why we did not take into account the in–the–money condition for the theoretical study. However, considering the set of martingales N^λ is far more efficient from a practical point of view.

In our numerical examples, we modify $V_{p,n}$ and $V_{p,n}^m$ to take into account this improvement and consider instead

$$\tilde{V}_{p,n}(\lambda) = \mathbb{E}\left[\max_{\tau_0\le k\le n}(Z_{t_k} - N_k(\lambda))\right] \quad \text{and} \quad \tilde{V}_{p,n}^m(\lambda) = \frac{1}{m}\sum_{i=1}^{m}\max_{\tau_0\le k\le n}(Z_{t_k}^{(i)} - N_k^{(i)}(\lambda)).$$

The idea of using martingales starting from the first time the option goes in the money is actually owed to [80]. Although he did not discuss it much, this was his choice in the examples he treated.

5.4.2 Our implementation of the algorithm

To practically compute the infimum of $\tilde{V}_{p,n}^m$, we advise to use a gradient descent algorithm, see Algorithm 5.4.1. The efficiency of such an approach mainly depends on the computation of the descend direction. When the problem is not twice differentiable, the gradient at the current point is used as a descent direction but it often needs to be scaled, which makes the choice of the step size α_ℓ a burning issue to ensure a fast numerical convergence. We refer to [20] for a comprehensive survey of several step size rules. After many tests, we found that the step size rule proposed by [76] was the best performing one in our context

$$\alpha_\ell = \frac{\tilde{V}_{p,n}^m(x_\ell) - v^\sharp}{\left\|\nabla\tilde{V}_{p,n}^m(x_\ell)\right\|^2}$$

where $v^\sharp$ is the price of the American option we are looking for. In practice, we use the price of the associated European option instead of $v^\sharp$, which makes α_ℓ too large and explains the need of the magnitude factor γ in Algorithm 5.4.2. The value of the European price does not need to be very accurate. A decent and fast approximation can be computed with a few thousand samples within few seconds no matter the dimension of the problem.

To better understand how this algorithm works, it is important to note that as $N(\lambda)$ linearly depends on λ, $N(\lambda) = \lambda\cdot\nabla_\lambda N(\lambda)$ and therefore both the value function and its gradient are computed at the same time without extra cost. So, $\nabla\tilde{V}_{p,n}^m(x_{\ell+1})$ is not actually computed on line 9 but at the same time as $v_{\ell+1/2}$ on line 5.

1 Generate $(G^{(1)}, Z^{(1)}), \dots, (G^{(m)}, Z^{(m)})$ m i.i.d. samples following the law of (Z, G)
2 $x_0 \leftarrow 0 \in \mathbb{R}^{A_{p,n}^{\otimes d}}$
3 $\ell \leftarrow 0$, $\gamma \leftarrow 1$, $d_0 \leftarrow 0$, $v_0 \leftarrow \infty$
4 **while** *True* **do**
5 Compute $v_{\ell+1/2} \leftarrow \tilde{V}_{p,n}^m(x_\ell - \gamma\alpha_\ell d_\ell)$
6 **if** $v_{\ell+1/2} < v_\ell$ **then**
7 $x_{\ell+1} \leftarrow x_\ell - \gamma\alpha_\ell d_\ell$
8 $v_{\ell+1} \leftarrow v_{\ell+1/2}$
9 $d_{\ell+1} \leftarrow \nabla\tilde{V}_{p,n}^m(x_{\ell+1})$
10 **if** $\frac{|v_{\ell+1}-v_\ell|}{v_\ell} \leq \varepsilon$ **then return**
11 **else**
12 $\gamma \leftarrow \gamma/2$
13 **end**
14 **end**

Algorithm 5.4.1: Sample Average Approximation of the dual price

The HPC approach. Our method targets large problems with as many as several thousands of components for λ. This requires to design a scalable algorithm capable of making the most of cluster architectures with hundreds of nodes. At each iteration, the computation of $\tilde{V}_{p,n}^m$ and $\nabla\tilde{V}_{p,n}^m$ is nothing but a standard Monte Carlo method and it inherits from its embarrassingly parallel nature.

A parallel algorithm for distributed memory systems based on the master/slave paradigm is proposed in Algorithm 5.4.2. At the beginning, each process samples a bunch of the m paths (lines 1–3). Then, at each iteration the master process broadcasts the values of d_ℓ, x_ℓ, α_ℓ and γ (line 7 of Algorithm 5.4.1). With these new values, each process computes its contribution to $\tilde{V}_{p,n}^m(x_\ell - \gamma\alpha_\ell d_\ell)$ and $\nabla\tilde{V}_{p,n}^m(x_\ell - \gamma\alpha_\ell d_\ell)$ (lines 8–9) and the Monte Carlo summations are obtained by two simple reductions (line 11). Then, the master process tests whether the move is admissible and updates the parameter for the next iteration or returns the solution if the algorithm is not moving enough anymore. This part carried out by the master process is very fast compared to the rest of the code and we dare say that there is no centralized computation in our algorithm. Moreover the communications are reduced to fours broadcasts, which guarantees an almost perfect very good scalability. The number of communications is monitored by the number of function evaluations, which remains quite small (between 10 and 20). We study the efficiency of our algorithm on a few examples at the end of Section 5.5.

Study of the complexity. Most of the computational time is spent computing the martingale part; remember that the cardinality of $\mathcal{C}_{p,n}$ is given by $\binom{nd+p}{nd} = \frac{(nd+p)\dots(nd+1)}{p!}$. Using martingales only starting once the option has been in the money enables us to only compute the martingale part on paths going in the money strictly before maturity time. Depending on the product, this may allow for a lot of computational time savings. The complexity of one iteration of the loop line 3 in Algorithm 5.4.1 is proportional to

$$\sharp\{\text{paths in the money strictly before time T}\} \times \binom{nd+p}{nd}.$$

The payoffs are computed once and for all before starting the descent algorithm. It is worth noting that its computational cost becomes negligible compared to the optimization part when the dimension

In parallel do
  Generate $(G^{(1)}, Z^{(1)}), \dots, (G^{(m)}, Z^{(m)})$ m i.i.d. samples following the law of (Z, G)
end
$x_0 \leftarrow 0 \in \mathbb{R}^{A_{p,n}^{\otimes d}}$
$\ell \leftarrow 0$, $\gamma \leftarrow 1$, $d_0 \leftarrow 0$, $v_0 \leftarrow \infty$
while *True* **do**
  Broadcast x_ℓ, d_ℓ, γ, α_ℓ
  In parallel do
    Compute $\max_{\tau_0 \le k \le n}(Z_{t_k}^{(i)} - N_k^{(i)}(x_\ell - \gamma\alpha_\ell d_\ell))$ for $i = 1, \dots, m$
  end
  Make a reduction of *the above contributions to obtain* $\tilde{V}_{p,n}^m(x_\ell - \gamma\alpha_\ell d_\ell)$ *and* $\nabla\tilde{V}_{p,n}^m(x_\ell - \gamma\alpha_\ell d_\ell)$
  $v_{\ell+1/2} \leftarrow \tilde{V}_{p,n}^m(x_\ell - \gamma\alpha_\ell d_\ell)$
  if $v_{\ell+1/2} < v_\ell$ **then**
    $x_{\ell+1} \leftarrow x_\ell - \gamma\alpha_\ell d_\ell$
    $v_{\ell+1} \leftarrow v_{\ell+1/2}$
    $d_{\ell+1} \leftarrow \nabla\tilde{V}_{p,n}^m(x_{\ell+1})$
    if $\frac{|v_{\ell+1} - v_\ell|}{v_\ell} \le \varepsilon$ **then return**
  else
    $\gamma \leftarrow \gamma/2$
  end
end

Algorithm 5.4.2: Parallel implementation of the Sample Average Approximation of the dual price

of the model or the number of dates increase, the most demanding computation being the evaluation of the martingale decomposition.

5.5 Applications

5.5.1 Some frameworks satisfying the assumption of Proposition 5.3.4

Let $(r_t)_t$ be the instantaneous interest rate supposed to be deterministic.

5.5.1.1 A put basket option in the multi–dimensional Black Scholes model

The $d-$dimensional Black Scholes model writes for $j \in \{1, \dots, d\}$

$$dS_t^j = S_t^j((r_t - \delta^j)dt + \sigma^j L_j dB_t)$$

where B is a Brownian motion with values in $\mathbb{R}^d$, $\sigma_t = (\sigma_t^1, \dots, \sigma_t^d)$ is the vector of volatilities, assumed to be deterministic and positive at all times, $\delta = (\delta^1, \dots, \delta^d)$ is the vector of instantaneous dividend rates and L_j is the j-th row of the matrix L defined as a square root of the correlation matrix Γ, ie. $\Gamma = LL'$. Moreover, we assume that L is lower triangular. Clearly, for every t, the random vector S_t is an element of $\mathbb{D}^{1,2}$.

The payoff of the put basket option writes as $\phi(S_t) = \left(K - \sum_{i=1}^d \omega^j S_t^j\right)_+$ where $\omega = (\omega^1, \ldots, \omega^d)$ is a vector of real valued weights. The function ϕ is Lipschitz continuous and hence $\phi(S_t) \in \mathbb{D}^{1,2}$ for all t. Moreover, for $s \leq t$ and $q \in \{1, \ldots, d\}$, we have on the set $\{\phi(S_t) > 0\}$

$$D_s^q \phi(S_t) = \sum_{j=1}^d \omega^j S_t^j \sigma^j L_{j,q}.$$

In particular for $q = d$, we get $D_s^d \phi(S_t) = \omega^d S_t^d \sigma^d L_{d,d}$.

Let $1 \leq k \leq n$ and F be a non zero and $\mathcal{F}_{t_k}-$measurable element of $\mathcal{C}_{p-1,n}$, ie.

$$F = \sum_{\alpha \in A_{p-1,n}^{\otimes d,k}} \lambda_\alpha \widehat{H}_\alpha^{\otimes d}(G_1, \ldots, G_n)$$

for some $\lambda \in \mathbb{R}^{A_{p,n}^{\otimes d}}$. Let $1 \leq r \leq k$.

$$\begin{aligned}
&\mathbb{P}\left(\forall t \in]t_{r-1}, t_r],\ D_t^d \phi(S_{t_k}) + F = 0 \mid \phi(S_{t_k}) > 0\right) \\
&= \mathbb{P}\left(\forall t \in]t_{r-1}, t_r],\ \omega^d S_{t_k}^d \sigma_t^d L_{d,d} + F = 0 \mid \phi(S_{t_k}) > 0\right) \\
&\leq \frac{\mathbb{P}\left(\forall t \in]t_{r-1}, t_r],\ \omega^d S_{t_k}^d \sigma_t^d L_{d,d} + F = 0\right)}{\mathbb{P}(\phi(S_{t_k}) > 0)}.
\end{aligned} \tag{5.20}$$

If $p = 1$, then F is a deterministic non zero constant. In this case, the numerator vanishes because $S_{t_k}^d$ has a density. Assume $p \geq 2$, then F is a multivariate polynomial with global degree $p - 1 \geq 1$. Then we can find $\ell \in \{1, \ldots, k\}$, $q \in \{1, \ldots, d\}$ and α such that $\alpha_\ell^q \geq 1$ and $\lambda_\alpha \neq 0$. Let $\hat{\mathcal{G}}$ be the sigma algebra generated by $(G_i^j, 1 \leq i \leq k, 1 \leq j \leq d, (i,j) \neq (\ell, q))$.

$$\mathbb{P}\left(\forall t \in]t_{r-1}, t_r],\ \omega^d S_t^d \sigma_t^d L_{d,d} + F = 0\right) = \mathbb{E}\left[\mathbb{P}\left(\forall t \in]t_{r-1}, t_r],\ \omega^d S_t^d \sigma_t^d L_{d,d} + F = 0 \mid \hat{\mathcal{G}}\right)\right].$$

Conditioning on $\hat{\mathcal{G}}$, the random variable $\omega^d S_t^d \sigma_t^d L_{d,d} + F$ only depends on G_ℓ^q. Consider the algebraic equation for $x \in \mathbb{R}$

$$a\, \mathrm{e}^{bx+c} = P(x) \tag{5.21}$$

where $(a, b, c) \in \mathbb{R}^3, a \neq 0, b \neq 0$ and P is polynomial with degree $p - 1 \geq 1$. Let $f(x) = a\, \mathrm{e}^{bx+c} - P(x)$, $f^{(p)}(x) = ab^p\, \mathrm{e}^{bx+c}$. Clearly, $f^{(p)}$ never vanishes, which ensures that f has at most p different roots. Hence, we deduce that for any $t \in]t_{r-1}, t_r]$, $\mathbb{P}\left(\omega^d S_t^d \sigma_t^d L_{d,d} + F = 0 \mid \hat{\mathcal{G}}\right) = 0$. Combining this result along with (5.20) proves that Equation (5.19) holds in this setting.

5.5.1.2 A put option on the minimum of a basket in the multi–dimensional Black Scholes model

We use the notation of the previous example. The payoff of the put option on the minimum of d assets write $\phi(S_t) = (K - \min_j(S_t^j))_+$. One can prove by induction on d that the function $x \in \mathbb{R}^d \longmapsto \min_j(x^j)$ is $1-$Lipschitz for the $1-$norm on $\mathbb{R}^d$. Hence, as the positive part function is also Lipschitz, the payoff function ϕ is Lipschitz. Then, [75, Proposition 1.2.4] yields that for all $t \in [0, T]$, $\phi(S_t) \in \mathbb{D}^{1,2}$ and for all $q \in \{1, \ldots, d\}$,

$$D^q(\phi(S_t)) = \sum_{j=1}^d \partial_{x^j} \phi(S_t) D^q(S_t^j) = \sum_{j=1}^d \partial_{x^j} \phi(S_t) S_t^j \sigma^j L_{j,q}.$$

With our choice for the matrix L,

$$D^d(\phi(S_t)) = \partial_{x^d}\phi(S_t)S_t^d\sigma^d L_{d,d} = -S_t^d\sigma^d L_{d,d}\mathbf{1}_{\phi(S_t)>0}\mathbf{1}_{\min_j(S_t^j)=S_t^d}.$$

Let $1 \leq k \leq n$ and F be a non zero and $\mathcal{F}_{t_k}-$measurable element of $\mathcal{C}_{p-1,n}$. For $1 \leq r \leq k$,

$$\begin{aligned}
&\mathbb{P}\left(\forall t \in]t_{r-1}, t_r],\ D_t^d\phi(S_{t_k}) + F = 0 \mid \phi(S_{t_k}) > 0\right) \\
&= \mathbb{P}\left(\forall t \in]t_{r-1}, t_r],\ -S_t^d\sigma^d L_{d,d} + F = 0 \mid \phi(S_{t_k}) > 0,\ \min_j(S_t^j) = S_t^d\right)\mathbb{P}\left(\min_j(S_t^j) = S_t^d\right) \\
&\quad + \mathbb{P}\left(\forall t \in]t_{r-1}, t_r],\ F = 0 \mid \phi(S_{t_k}) > 0,\ \min_j(S_t^j) \neq S_t^d\right)\mathbb{P}\left(\min_j(S_t^j) \neq S_t^d\right)
\end{aligned}$$

Clearly, the second term in the above sum is zero as F has a density. Hence,

$$\mathbb{P}\left(\forall t \in]t_{r-1}, t_r],\ D_t^d\phi(S_{t_k}) + F = 0 \mid \phi(S_{t_k}) > 0\right) \leq \frac{\mathbb{P}\left(\forall t \in]t_{r-1}, t_r],\ -S_t^d\sigma^d L_{d,d} + F = 0\right)}{\mathbb{P}(\phi(S_{t_k}) > 0)}.$$

We conclude as in the case of the put basket option.

5.5.1.3 A put option in the Heston model

The Heston model can be written

$$\begin{aligned}
dS_t &= S_t(r_t dt + \sqrt{\sigma_t}(\rho dW_t^1 + \sqrt{1-\rho^2}dW_t^2) \\
d\sigma_t &= \kappa(\theta - \sigma_t)dt + \xi\sqrt{\sigma_t}dW_t^1.
\end{aligned}$$

For $s \leq t$, $D_s^2 S_t = S_t\sqrt{1-\rho^2}\sqrt{\sigma_t}$. Conditionally on W^1, $D_s^2 S_t$ writes as $a\,\mathrm{e}^{bW_t^2+c}$ and we can unfold the same reasoning as after (5.21).

5.5.2 Numerical experiments

In this part, we present the results obtained from a sequential implementation of our approach as described in Algorithm 5.4.1. The computations are run on a standard laptop with an Intel Core i5 processor 2.9 Ghz. For each experiment, we report the price obtained using Algorithm 5.4.1 along with its computational time and standard deviation.

5.5.2.1 Examples in the Black Scholes models

We consider the $d-$dimensional Black Scholes as presented in Section 5.5.1.1. For the sake of simplicity in choosing the parameters, we have decide to use the same correlation between all the assets, which amounts to considering the following simple structure for Γ.

$$\Gamma = \begin{pmatrix} 1 & \rho & \dots & \rho \\ \rho & 1 & \ddots & \vdots \\ \vdots & \ddots & \ddots & \rho \\ \rho & \dots & \rho & 1 \end{pmatrix}$$

where $\rho \in]-1/(d-1), 1]$ to ensure that Γ is positive definite.

A basket option in the Black–Scholes model. We consider a put option on several assets as presented in Section 5.5.1.1. We report in Table 5.1 the price obtained with our approach for $m = 20,000$. The last column *reference price* corresponds to the prices reported in [85] on the same examples. These *reference* prices were obtained within a few minutes according to the authors whereas here we manage to get similar values within a few seconds. We can see that a second order chaos expansion, $p = 2$, already gives very accurate results within a few tenths of a second for a $5-$dimensional problem with 6 dates, which proves the impressive efficiency of our approach.

p	n	S_0	price	Stdev	time (sec.)	reference price
2	3	100	2.27	0.029	0.17	2.17
3	3	100	2.23	0.025	0.9	2.17
2	3	110	0.56	0.014	0.07	0.55
3	3	110	0.53	0.012	0.048	0.55
2	6	100	2.62	0.021	0.91	2.43
3	6	100	2.42	0.021	14	2.43
2	6	110	0.61	0.012	0.33	0.61
3	6	110	0.55	0.008	10	0.61

Table 5.1: Prices for the put basket option with parameters $T = 3$, $r = 0.05$, $K = 100$, $\rho = 0$, $\sigma^j = 0.2$, $\delta^j = 0$, $d = 5$, $\omega^j = 1/d$.

A call on the maximum of d assets in the Black–Scholes model. We consider a call option on the maximum of d assets in the Black Scholes model. As in the previous example, the last column *reference price* corresponds to the prices reported in [85] on the same examples. With no surprise,

d	p	m	S_0	price	Stdev	time (sec.)	reference price
2	2	20,000	90	10.18	0.07	0.4	8.15
2	3	20,000	90	8.5	0.05	4.1	8.15
2	2	20,000	100	16.2	0.06	0.54	14.01
2	3	20,000	100	14.4	0.06	5.6	14.01
5	2	20,000	90	21.2	0.09	2	16.77
5	3	40,000	90	16.3	0.05	210	16.77
5	2	20,000	100	30.7	0.09	3.4	26.34
5	3	40,000	100	26.0	0.05	207	26.34

Table 5.2: Prices for the call option on the maximum of d assets with parameters $T = 3$, $r = 0.05$, $K = 100$, $\rho = 0$, $\sigma^j = 0.2$, $\delta^j = 0.1$, $n = 9$.

the computational time increases exponentially with the dimension $n \times d$ and the degree p. Whereas a second order expansion provides very accurate results for the basket option, it only gives a rough upper–bound for the call option on the maximum of d assets. Considering a third order expansion $p = 3$ takes far longer but enables us to get very tight upper–bounds.

A geometric basket option in the Black–Scholes model Benchmarking a new method on high dimensional products becomes hardly feasible as almost no high dimensional American options can be priced accurately in a reasonable time. An exception to this is the geometric option with payoff $(K - (\prod_{j=1}^{d} S_t^j)^{1/d})_+$ for the put option. Easy calculations show that the price of this $d-$dimensional

option equals the one of the $1-$dimensional option with parameters

$$\hat{S}_0 = \left(\prod_{j=1}^{d} S_0^j\right)^{1/d} \quad ; \quad \hat{\sigma} = \frac{1}{d}\sqrt{\sum_{i,j}\sigma^i\sigma^j\Gamma_{ij}}; \quad \hat{\delta} = \frac{1}{d}\sum_{j=1}^{d}\left(\delta^j + \frac{1}{2}(\sigma^j)^2\right) - \frac{1}{2}(\hat{\sigma})^2.$$

Table 5.3 summarizes the correspondence values used in the examples.

d	S_0	σ	ρ	$\hat{S}_0$	$\hat{\sigma}$	$\hat{\delta}$
2	100	0.2	0	100	0.14	0.01
10	100	0.3	0.1	100	0.131	0.036
40	100	0.3	0.1	100	0.105	0.039

Table 5.3: Correspondence table for the parameters of the geometric options with $\delta^j = 0$.

d	σ^j	ρ	p	m	price	Stdev	time(sec)	$1-$d price
2	0.2	0	2	5000	4.32	0.04	0.018	4.20
2	0.2	0	3	5000	4.15	0.04	1.3	4.20
10	0.3	0.1	1	5000	5.50	0.06	0.12	4.60
10	0.3	0.1	2	20000	4.55	0.02	17	4.60
40	0.3	0.1	1	10000	4.4	0.03	1.4	3.69
40	0.3	0.1	2	20000	3.61	0.02	170	3.69

Table 5.4: Prices for the geometric basket put option with parameters $T = 1$, $r = 0.0488$ (it corresponds to a 5% annual interest rate), $K = 100$, $\delta^j = 0$, $n = 9$.

The $1 - d$ price is computed using a tree method with several thousand steps. We can see in Table 5.4 that a second order approximation gives very accurate result within a few seconds for an option with 10 underlying assets, which proves the efficiency of our approach. We cannot beat the curse of dimensionality, which slows down our algorithm for very large problems. For an option on 40 assets, we obtain a price up to a 3% relative error within 3 minutes which is already very fast for such a high dimensional problem. The number of terms involved in the chaos expansion can become very large: for $d = 40$ and $p = 2$, there are 65340 elements in $\mathcal{C}_{p,n}$. Even though we are not working in a linear algebra framework, it is advisable to ensure that the number of samples m used in the sample average approximation is larger than the number of free parameters in the optimization problem. When m becomes too small, we may face an over–fitting phenomenon as the number of parameters is far too large compared to the information contained in the sample average approximation. This probably explains why the price obtained for $p = 2$, $d = 40$ and $m = 40$ is slightly smaller than the true price.

In the next paragraph, we test the scalability of Algorithm 5.4.2 on this particular examples for a larger number of samples.

5.5.2.2 Scalability of the parallel algorithm

We consider the $40-$dimensional geometric put option studied in Table 5.4 with $p = 2$ and test the scalability of our parallel implementation for $m = 200,000$. The tests are run on a BullX DLC supercomputer containing 190 nodes for a total of 3204 CPU cores. We report in Table 5.5 the results of our scalability study using from 1 to 512 cores. Despite the two levels of parallelism available on this

supercomputer, we have used a pure MPI implementation without any reference to multithread programming. We might improve the efficiency a bit using nested parallelism, but the results are already convincing enough and do not justify the need of a two level approach, which makes the implementation more delicate. The sequential Algorithm runs within one hour and a quarter whereas using 512 cores we manage to get the computational time down to a dozen of seconds, which corresponds to a 0.6 efficiency. Considering the so short wall time required by the run on 512 cores, keeping the efficiency at this level represents a great achievement. Note that with 128 cores, the code runs within a minute with an efficiency of three quarters. These experiments prove the impressive scalability of our algorithm.

#processes	time (sec.)	efficiency
1	4365	1
2	2481	0.99
4	1362	0.90
16	282	0.84
32	272	0.75
64	87	0.78
128	52	0.73
256	34	0.69
512	10.7	0.59

Table 5.5: Scalability of Algorithm 5.4.2 on the $40-$dimensional geometric put option described above with $T = 1$, $r = 0.0488$, $K = 100$, $\sigma^j = 0.3$, $\rho = 0.1$, $\delta^j = 0$, $n = 9$, $p = 2$.

5.6 Conclusion

We have proposed a purely dual algorithm to compute the price of American or Bermudan options using some stochastic optimization tools. The starting point of our algorithm is the use of Wiener chaos expansion to build a finite dimensional vector space of martingales. Then, we rely on a sample average approximation to effectively optimize the coefficients of the expansion. Our algorithm is very fast: for problems up to dimension 5, a price is obtained within a few seconds, which is a tremendous improvement compared to existing purely dual methods. For higher dimensional problems, we can use a very scalable parallel algorithm to tackle very high dimensional problems (40 underlying assets). We can transparently deal with complex path–dependent payoffs without any extra computational cost. Event though, we restricted to a Brownian setting in this work, our approach could easily be extended to jump diffusion models by introducing Poisson chaos expansion, which is linked to Charlier polynomials (see [40]). We believe that our approach could be improved by cleverly reducing the number of terms in the chaos expansion, the computation of which centralizes most of the effort.

Chapter 6

Stochastic modelling of a ferromagnetic nano particle

This chapter based on [L-6, L-22] summarizes my contributions on the stochastic modelling of ferromagnetic nanoparticles, whose goal is to provide a mathematical framework to understand thermal effects. This is a joined work with Stéphane Labbé.

6.1 Introduction

The use of stochastic modelling for ferromagnetic particles goes back to the seminal paper [25] on the physical aspects of the problem. In this work, we focus on a single ferromagnetic mono–domain particle submitted to an external field, whose behaviour is usually modelled by the following deterministic dynamical system:

$$\frac{d\mu}{dt} = -\mu \wedge b - \alpha\mu \wedge (\mu \wedge b), \quad \mu_0 \in \mathcal{S}(\mathbb{R}^3) \tag{6.1}$$

where $b \in \mathbb{R}^3$ is the external magnetic field, $\alpha \in \mathbb{R}_+$ and $\mathcal{S}(\mathbb{R}^3)$ classically denotes the unit sphere in $\mathbb{R}^3$. It is clear that $|\mu_t| = 1$ for all $t \geq 0$. We introduce the antisymmetric operator $L : \mathbb{R}^3 \longmapsto \mathcal{M}_{3\times 3}$ associated to the cross product in $\mathbb{R}^3$

$$L(x) = \begin{pmatrix} 0 & -x^3 & x^2 \\ x^3 & 0 & -x^1 \\ -x^2 & x^1 & 0 \end{pmatrix}.$$

Let $A : \mathbb{R}^3 \longrightarrow \mathcal{M}_{3\times 3}$ be the operator defined by

$$A(x) = \alpha x^T x I - \alpha x x^T - L(x)$$

where I is the identity matrix in $\mathcal{M}_{3\times 3}$. Note that $x^T A(x) = 0$ for all $x \in \mathbb{R}^3$. We can write (6.1) as

$$\frac{d\mu}{dt} = A(\mu)b, \quad \mu_0 \in \mathcal{S}(\mathbb{R}^3). \tag{6.2}$$

This chapter aims at introducing stochastic perturbations in order to model thermal effects. In Section 6.2, we present three different ways of introducing a stochastic perturbation in (6.2) while preserving the property $|\mu_t| = 1$ for all $t \geq 0$. In Section 6.3, we study the long time behaviour of the stochastic model developed at the end of Section 6.2. Finally, we present some numerical simulations illustrating the theoretical results.

6.2 Stochastic modelling issues

Let $(\Omega, \mathcal{A}, \mathbb{F} = (\mathcal{F}_t)_{t\geq 0}, \mathbb{P})$ be a filtered probability space. We consider a standard $\mathbb{F}-$Brownian motion W with values in $\mathbb{R}^3$. Thermal effects can be embedded in the deterministic system (6.2) by adding a stochastic perturbation to the external field b, which naturally leads to the following stochastic system

$$d\mu_t = A(\mu_t)bdt + \varepsilon A(\mu_t)dW_t \tag{6.3}$$

where $\varepsilon > 0$ controls the magnitude of the stochastic perturbation. If this SDE is interpreted in the Itô sense, $d\left|\mu_t\right|^2 = \varepsilon^2 \operatorname{tr}(A(\mu_t)A^T(\mu_t))dt > 0$. Hence, the system is not physically consistent as it does not preserve $|\mu_t|$ whereas this is a physical invariant of (6.2), see [24]. In this section, we investigate several ways of modifying the stochastic model to ensure its consistency with the physical model.

6.2.1 Rescaling the Itô approach

In [L-6], we proposed to introduce a rescaling step in the above equation as we do when discretizing the ODE (6.2) with a non conservative scheme. Consider the coupled SDE

$$\begin{cases} dY_t &= A(\mu_t)bdt + \varepsilon A(\mu_t)dW_t, \quad Y_0 \in \mathcal{S}(\mathbb{R}^3) \\ \mu_t &= \frac{Y_t}{|Y_t|}. \end{cases} \tag{6.4}$$

We can prove that the process $(|Y_t|)_t$ is actually deterministic (see [L-6]).

Proposition 6.2.1 *Let (Y, μ) be a pair of processes satisfying* (6.4)*, then*

$$d\left|Y_t\right|^2 = 2\varepsilon^2(\alpha^2 + 1)dt$$

Let $h(t) = \sqrt{2\varepsilon^2(\alpha^2+1)t + 1}$. The process μ solves an autonomous equation

$$d\mu_t = \left(-\frac{h'(t)}{h(t)}\mu_t + \frac{A(\mu_t)}{h(t)}b\right)dt + \frac{\varepsilon}{h(t)}A(\mu_t)dW_t,$$

from which it clearly appears that the noise added to the model vanishes at the rate $\varepsilon h(t)^{-1}$, which looks quite arbitrary. The symmetry of the physical system suggests to focus on the component of μ_t along the vector b.

$$\begin{aligned} d(\mu_t \cdot b) = &\left\{-(\mu_t \cdot b)\frac{h'(t)}{h(t)} + \frac{\alpha}{h(t)}\left(|b|^2 - (\mu_t \cdot b)^2\right)\right\}dt \\ &- \frac{\varepsilon}{h(t)}\big(-L(\bar{\mu}_t)b + \alpha((\bar{\mu}_t \cdot b)\bar{\mu}_t - b)\big) \cdot dW_t. \end{aligned} \tag{6.5}$$

The asymptotic behaviour of this model was studied in detail in [L-6]. Here, we prefer to develop two other approaches which have been investigated more recently in [L-22] as we believe they are more versatile.

6.2.2 Pulling back the Itô approach

If instead of trying to introduce some rescaling mechanism in (6.3), we simply move the dynamics back on to the sphere with an extra term K, we get the following model

$$d\mu_t = A(\mu_t)bdt + \varepsilon A(\mu_t)dW_t + K_t dt.$$

Using that $x^T A(x) = 0$, we deduce that $d|\mu_t|^2 = 2\mu_t \cdot K_t dt + \mathrm{tr}(d\langle\mu\rangle_t)$. An easy computation shows that $\mathrm{tr}(d\langle\mu\rangle_t) = 2\varepsilon^2(\alpha^2|\mu_t|^2 + 1)|\mu_t|^2 dt$. Then, the condition $|\mu_t| = 1$ imposes to choose $K_t = -\varepsilon^2(\alpha^2|\mu_t|^2 + 1)\mu_t + K_t^\perp$, where $K_t^\perp$ is orthogonal to μ_t for all t. The final term K_t can be thought of as a pull back onto the sphere. The minimum norm pull is obtained by choosing $K^\perp = 0$, which leads to

$$d\mu_t = A(\mu_t)bdt + \varepsilon A(\mu_t)dW_t - \varepsilon^2(\alpha^2|\mu_t|^2 + 1)\mu_t dt$$

and it simplifies into

$$d\mu_t = (A(\mu_t)b - \varepsilon^2(\alpha^2 + 1)\mu_t)dt + \varepsilon A(\mu_t)dW_t \tag{6.6}$$

as $|\mu_t|^2 = 1$. This equation will show up later as the Itô form of the Stratonovich stochastic model. The idea of taking for the K_t term the Euclidean projection has been used to define the spherical Brownian motion, see [22].

6.2.3 The Stratonovich approach

As the Stratonovich stochastic calculus satisfy the standard differentiation rules, if we interpret (6.3) in the Stratonovich sense, the process μ will automatically satisfy $|\mu_t| = 1$ for all $t \geq 0$.

Let ∂ denote the Stratonovich differential operator as in [82]. Let $(\bar{\mu}_t)_t$ denote the stochastic system with a Stratonovich perturbation. We assume that the magnitude of the stochastic perturbation is given by a deterministic positive function $(\varepsilon_t)_t$. In this section, we consider the stochastic model defined by the following Stratonovich SDE

$$\partial\bar{\mu}_t = A(\bar{\mu}_t)b\partial t + \varepsilon_t A(\bar{\mu}_t)\partial W_t. \tag{6.7}$$

If we compute $\partial|\bar{\mu}_t|^2 = 2\bar{\mu}_t \cdot \partial\bar{\mu}_t$ using Equation (6.7), we immediately notice that $\partial|\bar{\mu}_t|^2 = 0$. Now, we turn this Stratonovich SDE into its Itô form (see [82, V.30])

$$d\bar{\mu}_t = (A(\bar{\mu}_t)b - \varepsilon_t^2(\alpha^2 + 1)\bar{\mu}_t)dt + \varepsilon A(\bar{\mu}_t)dW_t. \tag{6.8}$$

This equation is similar to (6.6), which was obtained by pulling the Itô process back onto the sphere. The similarity of the two equations actually advocates to interpret the noise in the Stratonovich sense as it naturally preserves the norm of μ. However, it is easy to check that

$$(d(\bar{\mu}_t \cdot b))\Big|_{\bar{\mu}_t = b/|b|} = -\varepsilon_t^2(\alpha^2 + 1)|b|\,dt.$$

This implies that b cannot be an equilibrium point of the stochastic system $(\bar{\mu}_t)_t$ unless ε_t goes to 0 for large t. It was proved in [L-6] that $\bar{\mu}_t$ could not converge to b. Actually, (6.8) very much looks like a continuous time stochastic approximation and it is known from [15, 39] that to obtain a long time stationary measure one has to let ε_t go to zero as well.

6.3 The Stratonovich model with decreasing noise

In this section, we assume that $(\varepsilon_t)_t$ is a decreasing function satisfying $\varepsilon_t > 0$ for all $t \geq 0$ and $\lim_{t\to\infty} \varepsilon_t = 0$.

$$d\bar{\mu}_t = -(A(\bar{\mu}_t)b + \varepsilon_t^2(\alpha^2 + 1)\bar{\mu}_t)dt + \varepsilon_t A(\bar{\mu}_t)dW_t,$$

where the operator A simplifies into $A(x) = \alpha I - \alpha xx^* - L(x)$ for $x \in \mathcal{S}(\mathbb{R}^3)$. Then, we deduce that

$$\begin{aligned} d(\bar{\mu}_t \cdot b) = &- \left\{ \alpha \left((\bar{\mu}_t \cdot b)^2 - |b|^2 \right) + \varepsilon_t^2 (\alpha^2 + 1) \bar{\mu}_t \cdot b \right\} dt \\ &- \varepsilon_t \big(- L(\bar{\mu}_t) b + \alpha ((\bar{\mu}_t \cdot b) \bar{\mu}_t - b) \big) \cdot dW_t. \end{aligned} \tag{6.9}$$

Remark 6.3.1 This model looks very much like (6.5) when taking $\varepsilon_t = \varepsilon / h(t)$ since $h'(t) = \varepsilon^2(\alpha^2 + 1)/h(t)$. The only difference is on the term $\alpha(|b|^2 - (\mu_t \cdot b)^2)$, which is divided by $h(t)$ in the rescaled Itô approach while it is not in (6.9). The dynamics of $(\bar{\mu}_t \cdot b)_t$ writes as the one obtained when taking the dot product between b and the deterministic system (6.1) plus an additive noise term.

6.3.1 The case $\alpha > 0$

Proposition 6.3.2 *Assume that one of the following conditions holds*

(i) $\int_0^\infty \varepsilon_t^2 dt = \infty$ *and* $\int_0^\infty \varepsilon_t^4 dt < \infty$.

(ii) $\int_0^\infty \varepsilon_t^2 dt < \infty$.

Then, $\bar{\mu}_t \xrightarrow[t \to \infty]{} \frac{b}{|b|}$ *a.s.*

The proof of this result heavily relies on a bespoke version of the ODE method, which aims at linking the behaviour of the SDE with the one of the ODE obtained by averaging the martingale part. A general theory for the ODE method was developed in different frameworks by [14, 16, 66]. For more results on the stability of SDE, we refer to [60]. The technique used in the second step of the proof below is very similar to the one developed in the proof of Theorem 2.2.1 in Chapter 2.

Sketch of the proof. As $|\bar{\mu}_t| = 1$, the result is equivalent to $\bar{\mu}_t \cdot b \to |b|$ a.s.
First, we prove that the martingale part of $(\bar{\mu}_t \cdot b)_t$ (see (6.9)) converges a.s. to some random variable M_∞. We set $M_t = \mathbb{E}[M_\infty | \mathcal{F}_t]$.

Second, define $X_t = \bar{\mu}_t \cdot b - (M_\infty - M_t)$. The process X solves the classical differential equation

$$dX_t = - \left\{ \alpha \left((\bar{\mu}_t \cdot b)^2 - |b|^2 \right) + \varepsilon_t^2 (\alpha^2 + 1) \bar{\mu}_t \cdot b \right\} dt. \tag{6.10}$$

Let $\eta > 0$, there exists $T > 0$ s.t. for all $t \geq T$, $\varepsilon_t^2(\alpha^2 + 1)\,|b| \leq \eta$ and $|\bar{\mu}_t \cdot b - X_t| \leq \eta$.

Let $0 < \delta_1 < \delta_2 < |b|$. We can choose η small enough such that $\delta_1 < \delta_2 - 2\eta$ and $\eta \leq \frac{\alpha}{2}(|b|^2 - \delta_2^2)$. Figure 6.3.1 defines three regions: the two pole caps and the region in between; depending on the position of $\bar{\mu}_t \cdot b$, we can bound from below the r.h.s of (6.10) to deduce that for $t \geq s > T$.

$$\begin{aligned} X_t - X_s \geq &\int_s^t (\alpha^2 + 1) \varepsilon_u^2 \delta_1 \mathbf{1}_{\bar{\mu}_u \cdot b \leq -\delta_1} du + \int_s^t \frac{\alpha}{2} (|b|^2 - \delta_2^2) \mathbf{1}_{-\delta_2 \leq \bar{\mu}_u \cdot b \leq \delta_2} du \\ &- \int_s^t |b|\, \varepsilon_u^2 (\alpha^2 + 1) \mathbf{1}_{\bar{\mu}_u \cdot b > \delta_2} du, \\ X_t - X_s \geq &\int_s^t (\alpha^2 + 1) \varepsilon_u^2 \delta_1 \mathbf{1}_{X_u \leq -\delta_1 - \eta} du + \int_s^t \frac{\alpha}{2} (|b|^2 - \delta_2^2) \mathbf{1}_{-\delta_2 + \eta \leq X_u \leq \delta_2 - \eta} du \\ &- \int_s^t |b|\, \varepsilon_u^2 (\alpha^2 + 1) \mathbf{1}_{X_u > \delta_2 + \eta} du. \end{aligned}$$

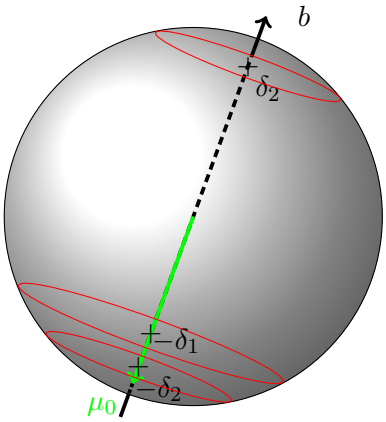


Figure 6.3.1: We consider three zones on the sphere $\{\bar{\mu} \cdot b \leq -\delta_1\}$, $\{-\delta_2 \leq \bar{\mu} \cdot b \leq \delta_2\}$ and $\{\bar{\mu} \cdot b \geq \delta_2\}$.

Note that X_t is increasing on the set $\{u \ : \ X_u \leq \delta_2 - \eta\}$. We can choose δ_2 sufficiently close to $|b|$ such that there exists t_1 for which $\int_s^{t_1} (\alpha^2+1)\varepsilon_u^2 \delta_1 \, du = |b| - (\delta_2 - \eta)$ — remember that η can be chosen as small as necessary. Hence, for all $t \geq t_1$, $X_t \geq -\delta_2 + \eta$. Therefore,

$$X_t - X_{t_1} \geq \int_{t_1}^t \frac{\alpha}{2}(|b|^2 - \delta_2^2)\mathbf{1}_{-\delta_2+\eta \leq X_u \leq \delta_2 - \eta} du - \int_{t_1}^t |b| \, \varepsilon_u^2(\alpha^2+1)\mathbf{1}_{X_u \geq \delta_2 + \eta} du.$$

From the continuity of X, we deduce that there exists $t_2 \geq t_1$ such that for all $t \geq t_2$, $X_t \geq \delta_2 - \eta$, which implies that for all $t \geq t_2$, $\bar{\mu}_t \cdot b \geq \delta_2 - 2\eta$. By choosing δ close to $|b|$ and η close to 0, we conclude that $\bar{\mu}_t \cdot b \to |b|$. ■

Proposition 6.3.3 *Assume $\bar{\mu}_t \to b/|b|$ a.s. If $(\varepsilon_t)_t$ is of class C^1 and $\lim_{t\to\infty} \frac{\varepsilon_t'}{\varepsilon_t} = 0$, then, for all $p \in \mathbb{N}$,*

$$\lim_{t\to\infty} \mathbb{E}\left[\left|\frac{b}{|b|} - \bar{\mu}_t\right|^{2p} \varepsilon_t^{-2p}\right] = \left(\frac{\alpha^2+1}{\alpha\,|b|}\right)^p p!.$$

Note that as $|\bar{\mu}_t| = 1$, $\left|\frac{b}{|b|} - \bar{\mu}_t\right|^2 = \frac{2}{|b|}(|b| - \bar{\mu}_t \cdot b)$. Hence, the proposition could equivalently write $\lim_{t\to\infty} \mathbb{E}[(|b| - \bar{\mu}_t \cdot b)^p \varepsilon_t^{-2p}] = \left(\frac{\alpha^2+1}{2\alpha}\right)^p p!$. For $p = 1$, this boils down to $\lim_{t\to\infty} \mathbb{E}[(|b| - \bar{\mu}_t \cdot b)\varepsilon_t^{-2}] = \frac{\alpha^2+1}{2\alpha}$. In the case of the rescaled Itô model, the convergence rate was given by $h(t)$, which actually monitors the magnitude of the noise. In the Stratonovich model, this role is played by ε_t. Hence, we would have expected a convergence rate of ε_t^{-1} whereas we obtained a much faster one, namely the square of it — ε_t^{-2}. Although in both models, the magnetic moment converges to b, the rates governing the convergence significantly differ.

Proposition 6.3.4 *Assume that*

- *there exists $\gamma > 0$ such that $\int_0^\infty \varepsilon_t^\gamma dt < \infty$;*
- *the function $(\varepsilon_t)_t$ is C^1, decreasing for large enough t and satisfies $\lim_{t\to\infty} \frac{\varepsilon_t'}{\varepsilon_t} = 0$.*

Then,

$$\text{for all } \eta > 0,\ \left|\frac{b}{|b|} - \bar{\mu}_t\right|^2 \varepsilon_t^{-2+\eta} \to 0\ a.s.$$

Sketch of the proof. **Step 1.** Once an L^p convergence rate has been established (see Proposition 6.3.3), such an almost sure result is easily deduced from Borrel Cantelli's Lemma for sequences indexed by a countable set and we obtain that

$$\lim_{t\in\mathbb{N}, t\to\infty} \left|\frac{b}{|b|} - \bar{\mu}_t\right|^2 \varepsilon_t^{-2+\eta} = 0\ a.s. \tag{6.11}$$

Extending this result to $t \in \mathbb{R}_+$ requires to monitor the behaviour in $L^p(\Omega)$ of $\left|\frac{b}{|b|} - \bar{\mu}_t\right|^2 \varepsilon_t^{-2}$ for $t \in [n, n+1]$ for any $n \in \mathbb{N}$.

Step 2. We aim at proving that $\lim_{n\to\infty} \sup_{n\le t\le n+1} \left| \left|\frac{b}{|b|} - \bar{\mu}_t\right|^2 \varepsilon_t^{-2+\eta} - \left|\frac{b}{|b|} - \bar{\mu}_n\right|^2 \varepsilon_n^{-2+\eta} \right| \to 0$ a.s. Define $X_t = \frac{b}{|b|} - \bar{\mu}_t$. Let $n \in \mathbb{N}$ and $n \le t \le n+1$

$$\begin{aligned}\left||X_t|^2 \varepsilon_t^{-2+\eta} - |X_n|^2 \varepsilon_n^{-2+\eta}\right| &\le C\left||X_n|^2 (\varepsilon_n^{-2+\eta} - \varepsilon_t^{-2+\eta})\right| + C\left|(|X_n|^2 - |X_t|^2)\varepsilon_t^{-2+\eta}\right| \\ &\le C\,|X_n|^2 \varepsilon_n^{-2+\eta} + 4C\,|X_n - X_t|\,\varepsilon_n^{-2+\eta}.\end{aligned}$$

As we know from (6.11) that $|X_n|^2 \varepsilon_n^{-2+\eta} \to 0$, it is sufficient to monitor $\sup_{n\le t\le n+1} |X_n - X_t|\,\varepsilon_n^{-2+\eta}$. Let $p > 1$. We deduce from Burkholder–Davis–Gundi's inequality that

$$\begin{aligned}\mathbb{E}\left[\sup_{n\le t\le n+1} |X_t - X_n|^{2p}\right] \le C\mathbb{E}\left[\int_n^{n+1} \alpha^{2p}\left(|b|^2 - (\bar{\mu}_u\cdot b)^2\right)^{2p} + \varepsilon_u^{4p}(\alpha^2+1)^{2p}\,du\right] \\ + C\mathbb{E}\left[\left(\int_n^{n+1} \varepsilon_u^2(\alpha^2+1)(|b|^2 - (\bar{\mu}_t\cdot b)^2)du\right)^p\right].\end{aligned}$$

From Proposition 6.3.3, $\mathbb{E}[(|b|^2 - (\bar{\mu}_u\cdot b)^2)^p] = O(\varepsilon_u^{2p})$ and we deduce that

$$\mathbb{E}\left[\sup_{n\le t\le n+1} \left(\varepsilon_n^{-2+\eta}\,|X_t - X_n|\right)^{2p}\right] \le C\varepsilon_n^{p\eta}.$$

For $p\eta \ge \gamma$, $\int_0^\infty \varepsilon_u^{p\eta} du < \infty$ and once more Borel–Cantelli's lemma yields that

$$\lim_{n\to\infty} \sup_{n\le t\le n+1} \varepsilon_n^{-2+\eta}\,|X_t - X_n|^2 = 0.$$

Then, we easily conclude that (6.11) holds for any real t and not only integers. ■

6.3.2 The case $\alpha = 0$

In this case, the process $\bar{\mu}$ solves the simplified equation

$$d\bar{\mu}_t = \left(L(b) - \varepsilon_t^2 I\right)\bar{\mu}_t dt - \varepsilon_t L(\bar{\mu}_t) dW_t. \tag{6.12}$$

We can integrate this SDE as a classical ODE to obtain

$$d\left(\mathrm{e}^{-L(b)t+\int_0^t \varepsilon_u^2 du}\,\bar{\mu}_t\right) = -\,\mathrm{e}^{-L(b)t+\int_0^t \varepsilon_u^2 du}\,\varepsilon_t L(\bar{\mu}_t)dW_t$$

$$\bar{\mu}_t - \mathrm{e}^{L(b)t-\int_0^t \varepsilon_u^2 du}\,\bar{\mu}_0 = -\,\mathrm{e}^{L(b)t-\int_0^t \varepsilon_u^2 du}\int_0^t \mathrm{e}^{-L(b)s+\int_0^s \varepsilon_u^2 du}\,\varepsilon_s L(\bar{\mu}_s)dW_s. \tag{6.13}$$

Let us introduce the square integrable martingale N defined by

$$N_t = \int_0^t \mathrm{e}^{-L(b)s+\int_0^s \varepsilon_u^2 du}\,\varepsilon_s L(\bar{\mu}_s)dW_s.$$

Proposition 6.3.5 *The long time behaviour of $(\bar{\mu}_t)_t$ depends on the integrability of $(\varepsilon_t)_t$.*

- *If $\int_0^\infty \varepsilon_u^2 du = \infty$, $\mathbb{E}[\bar{\mu}_t] \to 0$ when $t \to \infty$.*
- *When $\int_0^\infty \varepsilon_u^2 du < \infty$, N_t converges a.s. to some random N_∞ and*

$$\lim_{t\to\infty} \bar{\mu}_t - \mathrm{e}^{L(b)t-\int_0^\infty \varepsilon_u^2 du}(\bar{\mu}_0 - N_\infty) = 0 \ a.s$$

and moreover for all $p \in \mathbb{N}$, there exists c_p, such that $\mathbb{E}[|N_t|^{2p}] \le c_p\left(\mathrm{e}^{2\int_0^t \varepsilon_s^2 ds}-1\right)^p$ for all $t \ge 0$.

When $\int_0^\infty \varepsilon_u^2 du < \infty$, $\sup_t \mathbb{E}[|N_t|^2] < \infty$ and then N converges a.s. to N_∞ and $N_t = \mathbb{E}[N_\infty|\mathcal{F}_t]$. Hence, we clearly have $\mathbb{E}[N_\infty] = 0$ and we obtain that $\lim_{t\to\infty} \mathbb{E}[\bar{\mu}_t] - \mathrm{e}^{L(b)t-\int_0^t \varepsilon_u^2 du}\,\bar{\mu}_0 = 0$ a.s. The term $\mathrm{e}^{L(b)t}$ makes $\bar{\mu}_t$ move on the ring with level $\mathrm{e}^{-\int_0^\infty \varepsilon_u^2 du}(\bar{\mu}_0 \cdot b - N_\infty \cdot b)$.

If $\varepsilon_t = \frac{\varepsilon}{(t+1)^{1/2+\eta}}$ where $\eta > 0$ and $\varepsilon > 0$, we can explicitly compute the upper–bound on $\mathbb{E}[|N_t|^{2p}]$

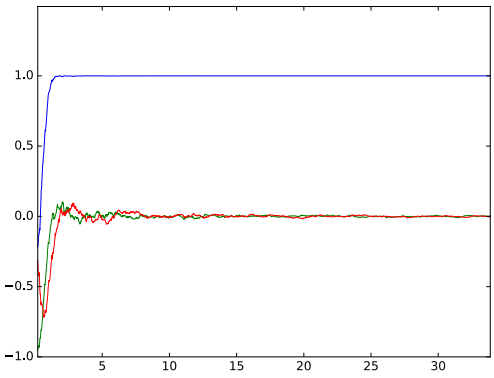
$$\left(\mathrm{e}^{2\int_0^t \varepsilon_s^2 ds}-1\right)^p = \left(\mathrm{e}^{\frac{\varepsilon^2}{\eta}(1-(t+1)^{-2\eta})}-1\right)^p \to \left(\mathrm{e}^{\frac{\varepsilon^2}{\eta}}-1\right)^p.$$

Usually, $\varepsilon^2 \ll \eta$ and therefore $\left(\mathrm{e}^{\frac{\varepsilon^2}{\eta}}-1\right)^p \approx \left(\frac{\varepsilon^2}{\eta}\right)^p$. Hence, with a very high probability, N_∞ remains tiny, and then for large t, $\bar{\mu}_t$ oscillates as $\mathrm{e}^{L(b)t-\int_0^\infty \varepsilon_u^2 du}\,\bar{\mu}_0$, which is non random.

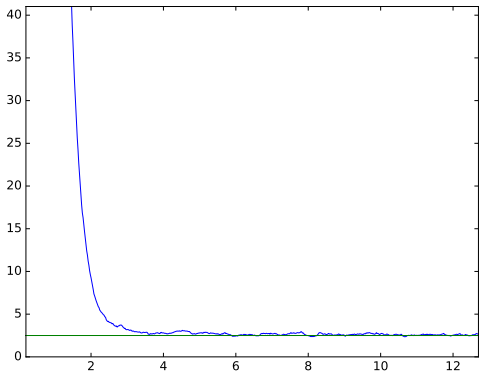
6.4 Numerical simulations

In this section, we illustrate the theoretical results of Section 6.3 for different values of α and functions $(\varepsilon_t)_t$ with different decaying rates. To discretize the Stratonovich model (6.7), we would rather consider its Itô form given by (6.8), on which we use an Euler scheme with time step Δt. The Euler scheme has the advantage of being fully explicit and is therefore easily to implement. We could have straightaway discretized the Stratonovich form (6.7), but the discretization of the Stratonovich integral must be performed using a semi–implicit scheme, which requires the use of a numerical solver at each iteration. In our case, the use of a semi–implicit scheme would have preserved the norm of the discretized process, which is not guarantied by an explicit scheme. However, using the Euler scheme on the Itô form did not raise any numerical difficulty.

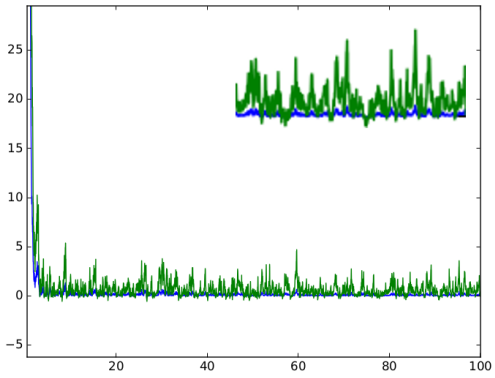
Some of the graphs below have required to compute expectations, which were approximated using a Monte Carlo method with 500 samples. This may seem few samples but it proved to be enough as the quantities involved have little variance especially when focusing on the behaviour for large times.



(a) A.s. convergence



(b) Convergence rate in L^2 of $\left|\frac{b}{|b|} - \bar{\mu}_t\right| \varepsilon_t^{-1}$



(c) Pathwise Convergence of $\left|\frac{b}{|b|} - \bar{\mu}_t\right| \varepsilon_t^{-1+\eta}$ for $\eta = 0.25$ (blue curve) or $\eta = 0.1$ (green curve)

Figure 6.4.1: Convergence of $(\bar{\mu}_t)_t$ for $\alpha = 2$, $\Delta t = 2 \times 10^{-2}$ and $\varepsilon_t = 0.1/(t+1)$.

6.4.1 The case $\alpha > 0$

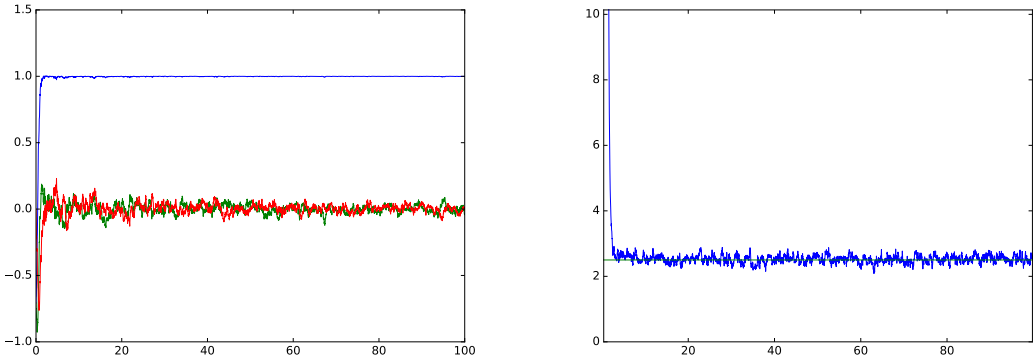
Figure 6.4.1 shows the convergence of $\bar{\mu}_t$ for ε_t satisfying $\int_0^\infty \varepsilon_t^2 dt < \infty$ when $\bar{\mu}_0$ is chosen such that $-1 < \bar{\mu}_0 \cdot b < 0$. The blue curve of Figure 6.1(a) corresponds to the component of $\bar{\mu}_t$ along the direction of b. We can see that the a.s. convergence of $\bar{\mu}_t$ to $b/|b|$ is very smooth and fast. We recover in Figure 6.1(b) the L^2 rate of Proposition 6.3.3. In particular, we notice that the transition phase is quite short as for $t = 10$ we already observe the numerical convergence. For the same parameters, Figure 6.1(c) illustrates the a.s. convergence rate result (see Proposition 6.3.4) for two values of η. Non surprisingly, the larger η, the smoother the convergence.

When the magnitude of the noise decays slowly, the convergence should be less smooth as suggested by Proposition 6.3.3, which corresponds to what we can see on Figure 6.4.2. Closely looking at Figures 6.1(a) and 6.2(a), we notice that the component of $\bar{\mu}$ along b converges faster that the two

others. Actually, from what we explained after Proposition 6.3.3, we have

$$\left|\frac{b}{|b|} - \bar{\mu}_t\right|^2 = \left(1 - \frac{b}{|b|} \cdot \bar{\mu}_t\right) + (\bar{\mu}_t \cdot e_2)^2 + (\bar{\mu}_t \cdot e_3)^2$$
$$(1 - \bar{\mu}_t \cdot e_1) = (\bar{\mu}_t \cdot e_2)^2 + (\bar{\mu}_t \cdot e_3)^2.$$

From this last equation, it is clear that there is a power 2 difference between the rates of convergence of $\bar{\mu}_t \cdot b$ and of the two other components, which fully matches our numerical observations.



(a) A.s convergence

(b) Convergence rate in L^1

Figure 6.4.2: A.s. convergence of $(\bar{\mu}_t)_t$ for $\alpha = 2$, $\Delta t = 2 \times 10^{-2}$ and $\varepsilon_t = 0.1/(t+1)^{1/3}$.

From the theoretical results of Section 6.3, it is clear that the noise term has a stabilizing effect on the system and is in particular responsible for escaping from $-b$, which is an unstable critical point of the deterministic system. Figure 6.4.3 confirms that the stabilizing effect exists even when the magnitude of the noise decays very fast — $(\varepsilon_t)_t$ belongs to $L^1([0,\infty))$ — and $\bar{\mu}_0 = -b/|b|$, which is the worst case scenario. After a very short transition period during which $\bar{\mu}$ circles around on the sphere while heading to $b/|b|$, the process stabilizes around its limit and remains impressively smooth.

6.4.2 The case $\alpha = 0$

As emphasized by the theoretical results, the behaviour of the process $(\bar{\mu}_t)_t$ depends very much on the value of α. When $\alpha = 0$ and there is no noise, $\bar{\mu}_t$ evolves on a circle with constant latitude. Actually, we recover a very similar behaviour in Figure 6.4.4 when the noise magnitude decays quickly — $\int_0^\infty \varepsilon_t dt < \infty$. Clearly, $\bar{\mu}_t$ heads to a constant latitude level and keeps turning on this parallel circle but unlikely to what happens in the deterministic case, the latitude is not exactly determined by $\bar{\mu}_0 \cdot b$ but is slightly randomly shifted as seen in Proposition 6.3.5. Closely looking at Figure 6.4.4, we can see that the magnitude of the oscillations tends to increase a little with time, which is a consequence of the discretized process not having a constant norm. This could be circumvented by considering a smaller discretization step Δt.

When the noise decreases slowly, ie. $\int_0^\infty \varepsilon_t^2 dt = \infty$, its effect remains over time and prevents any almost sure limiting behaviour to appear. The process $(\bar{\mu}_t)$ keeps wandering around on the sphere and we see from Figure 6.4.5 that $\mathbb{E}[\bar{\mu}_t] \to 0$.

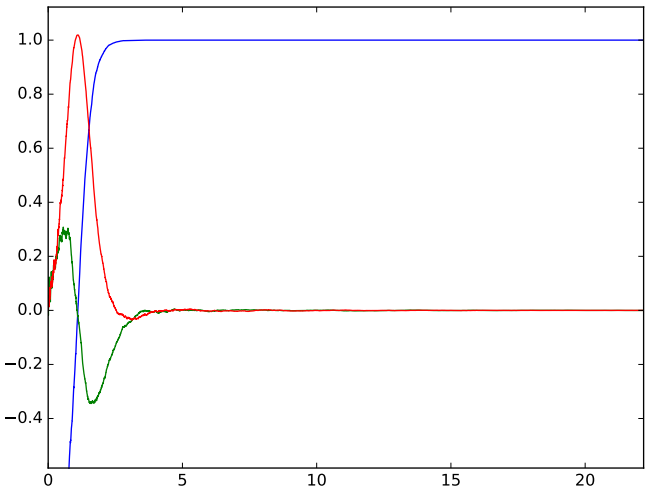


Figure 6.4.3: A.s. convergence of $(\bar{\mu}_t)_t$ for $\alpha = 2$, $\Delta t = 2{\times}10^{-3}$, $\bar{\mu}_0 = -b/\left|b\right|$ and $\varepsilon_t = 0.1/(t{+}1)^2$.

6.5 Conclusion

In this work, we have discussed issues on the stochastic modelling of a ferromagnetic nanoparticle. Among the different approaches, the Stratonovich approach with a decaying noise magnitude showed up as the most natural one. We investigated the long time behaviour of the model and proved its convergence to the unique stable equilibrium of the deterministic system when $\alpha > 0$. When $\alpha = 0$, the evolution of the system depends on the magnitude of the noise; when a limiting behaviour appears, the process keeps revolving on a parallel ring. All these theoretical results have been illustrated by numerical simulations, which help better understanding how thermal effects can be modelled in micromagnetism.

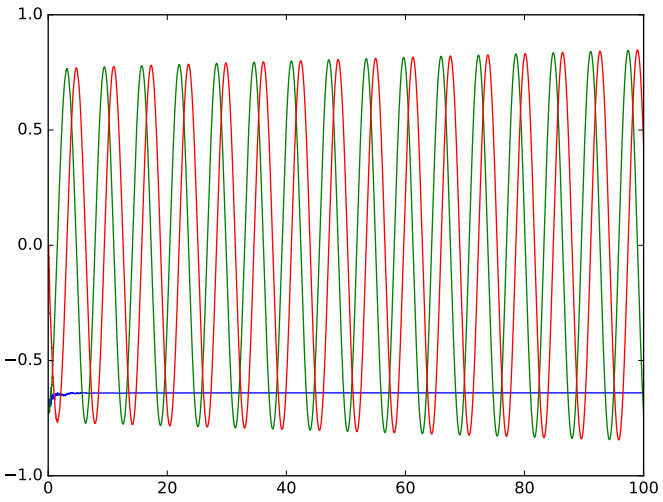


Figure 6.4.4: Convergence of $(\bar{\mu}_t)_t$ for $\varepsilon_t = 0.3/(t+1)^2$, $\Delta t = 2 \times 10^{-3}$.

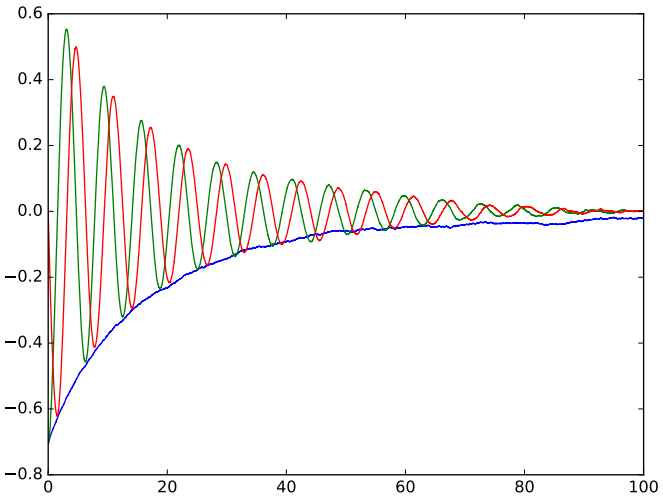


Figure 6.4.5: Convergence of $\mathbb{E}[\bar{\mu}_t]$ for $\varepsilon_t = 0.3/(t+1)^{0.1}$, $\Delta t = 2 \times 10^{-3}$.

Chapter 7

Some key technical tools

In this chapter, we gather some technical tools, which proved to be essential for getting some results presented in this manuscript and which we believe could be useful in other problems.

7.1 Strong law of large numbers for doubly indexed sequences

In this section, we state two corner stone results used at several places in this manuscript and proved in [L-1]. We tackle the convergence of empirical averages of doubly indexed random sequences when both indices tend to infinity together.

Proposition 7.1.1 *Let $(X_{n,m})_{n,m}$ be a sequence of vector valued random variables such that for all n, $\mathbb{E}[X_{n,m}] = x_m$ with $\lim_{m\to+\infty} x_m = x$. We define $\overline{X}_{n,m} = \frac{1}{n}\sum_{i=1}^{n} X_{i,m}$. Assume that the following two assumptions are satisfied*

(H7.1) *i.* $\sup_n \sup_m n \operatorname{Var}\left(\overline{X}_{n,m}\right) < +\infty$.

ii. $\sup_n \sup_m \operatorname{Var}\left(X_{n,m}\right) < +\infty$.

Then, for all increasing functions $\rho : \mathbb{N} \to \mathbb{N}$, $\overline{X}_{n,\rho(n)} \xrightarrow[n\to+\infty]{} x$ a.s. and in $\mathbb{L}^2$.

From this proposition, one can easily deduce the following corollary by extracting a bespoke subsequence.

Corollary 7.1.2 *Assume that $(X_{i,m})_{i,m}$ is a sequence of random vectors satisfying the assumptions of Proposition 7.1.1. Then, for any strictly increasing function $\xi : \mathbb{N} \to \mathbb{N}$, $\overline{X}_{\xi(n),n} \xrightarrow[n\to+\infty]{} x$ a.s. and in $\mathbb{L}^2$.*

Proof (Proof of Proposition 7.1.1). The proof of this result closely mimics the one of [78, Theorem IV.1.1]. We introduce the sequence $(Y_{i,m})_{i,m}$ defined by $Y_{i,m} = X_{i,m} - x_m$, which satisfies $\mathbb{E}[Y_{i,m}] = 0$. As $\lim_{m\to\infty} x_m = x$, it is sufficient to prove that $\overline{Y}_{n,\rho(n)} \xrightarrow[n\to+\infty]{} 0$ a.s.

Condition ($\mathcal{H}$7.1-*i)* implies the $\mathbb{L}^2$ convergence to 0. We introduce the sequence $(Z_{n,m})_n$ defined by $Z_{n,m} = \sup\{\left|\bar{Y}_{k,m}\right| \ : \ n^2 \le k < (n+1)^2\}$. Let k be such that $n^2 \le k < (n+1)^2$, then

$$\left|\bar{Y}_{k,m}\right| \le n^{-2}\left(n^2\left|\bar{Y}_{n^2,m}\right| + \sum_{i=n^2+1}^{k}|Y_{i,m}|\right),$$

$$Z_{n,m} \le \left|\bar{Y}_{n^2,m}\right| + \frac{1}{n^2}\sum_{i=n^2+1}^{(n+1)^2}|Y_{i,m}|.$$

Then,

$$\mathbb{E}[Z_{n,m}^2] \leq \mathbb{E}[\bar{Y}_{n^2,m}^2] + \sum_{i=n^2+1}^{(n+1)^2} \left(\frac{\mathbb{E}[|Y_{i,m}|^2]}{n^4} + 2\frac{\mathbb{E}[\left|\bar{Y}_{n^2,m}\right| |Y_{i,m}|]}{n^2} \right) + 2 \sum_{i,j=n^2+1; i\neq j}^{(n+1)^2} \frac{\mathbb{E}[|Y_{j,m}| \, |Y_{i,m}|]}{n^4}.$$

Let $\kappa > 0$ denote the maximum of the upper bounds involved in Assumption ($\mathcal{H}$7.1). Using the Cauchy Schwartz inequality, we get

$$\begin{aligned}\mathbb{E}[Z_{n,m}^2] &\leq \frac{\kappa}{n^2} + \frac{\kappa((n+1)^2 - n^2)}{n^4} + 2\frac{\kappa^2((n+1)^2 - n^2)}{n^3} + 2\frac{\kappa^2((n+1)^2 - n^2)^2}{n^4} \\ &\leq \frac{\kappa}{n^2} + \frac{\kappa(2n+1)}{n^4} + 2\frac{\kappa^2(2n+1)}{n^3} + 2\frac{\kappa^2(2n+1)^2}{n^4}.\end{aligned}$$

Hence, for any function $\rho : \mathbb{N} \to \mathbb{N}$, $\mathbb{E}[Z_{n,\rho(n)}^2] \leq Cn^{-2}$ where $C > 0$ is a constant independent of ρ. Therefore, we have $\mathbb{P}(Z_{n,\rho(n)} \geq n^{-1/4}) \leq Cn^{-3/2}$. This inequality implies using the Borel Cantelli Lemma that, for n large enough $Z_{n,\rho(n)} \leq n^{-1/4}$ a.s. which yields the a.s. convergence to 0. ■

Proposition 7.1.3 *Let $(F_{n,m})_{n,m}$ be a sequence of random variables with values in the set of continuous functions, ie. for all n, m, $F_{n,m} : \Omega \longrightarrow C^0(\mathbb{R}^d)$. Moreover, we assume that there exists a sequence of functions $(f_m)_m$ satisfying $\mathbb{E}[F_{n,m}] = f_m$ for all n. We define $\overline{F}_{n,m} = \frac{1}{n}\sum_{i=1}^n F_{i,m}$. Assume that the two following assumptions are satisfied*

($\mathcal{H}$7.2) *One of the following criteria holds*

- *i. The sequence $(f_m)_m$ converges pointwise to some continuous function f.*
- *ii. The sequence $(f_m)_m$ converges locally uniformly to some function f.*

($\mathcal{H}$7.3) *For any compact set $W \subset \mathbb{R}^d$,*

- *i.* $\sup_n \sup_m n \operatorname{Var}\left(\sup_{x\in W} \left|\overline{F}_{n,m}(x)\right|\right) < +\infty$.
- *ii.* $\sup_n \sup_m \operatorname{Var}\left(\sup_{x\in W} |F_{n,m}(x)|\right) < +\infty$.

($\mathcal{H}$7.4) *For all $y \in \mathbb{R}^d$,* $\lim_{\delta\to 0} \sup_n \sup_m \mathbb{E}\left[\sup_{|x-y|\leq\delta} |F_{n,m}(x) - F_{n,m}(y)|\right] = 0$.

Then, for all functions $\rho : \mathbb{N} \to \mathbb{N}$, the sequence of random functions $\overline{F}_{n,\rho(n)}$ converges a.s. locally uniformly to the locally continuous function f.

Remark 7.1.4 When for every fixed m, the sequence $(F_{n,m})_n$ is i.i.d., ($\mathcal{H}$7.4) is ensured by

$$\forall\, y \in \mathbb{R}^d,\ \lim_{\delta\to 0} \limsup_m \mathbb{E}\left[\sup_{|x-y|\leq\delta} |F_{1,m}(x) - F_{1,m}(y)|\right] = 0$$

and Assumption ($\mathcal{H}$7.3-*ii)* implies ($\mathcal{H}$7.3-*i)*.

As in Corollary 7.1.2, for any strictly increasing function $\xi : \mathbb{N} \to \mathbb{N}$, the sequence $\overline{F}_{\xi(n),n}$ converges a.s. locally uniformly to the locally continuous function f.

Proof. We can apply Proposition 7.1.1, to deduce that a.s. $\overline{F}_{n,\rho(n)}$ converges pointwise to the function f. If we do not already know that f is continuous, then thanks to ($\mathcal{H}$7.3-*ii)*, we can apply Lebesgue's theorem to deduce that the functions f_m are continuous. The uniform convergence of the sequence f_m to f (see ($\mathcal{H}$7.2-*ii)*) proves that the function f is continuous.

Let W be a compact set of $\mathbb{R}^d$, we can cover W with a finite number K of open balls W_k with centers $(x_k)_k$ and radiuses $(r_k)_k$, i.e. $W_k = B(x_k, r_k)$ and $W = \cup_{k=1}^K W_k$. We want to prove that $\sup_{x\in W} \left|\overline{F}_{n,\rho(n)}(x) - f(x)\right| \xrightarrow[n\to+\infty]{a.s.} 0$. We write

$$\sup_{x\in W} \left|\overline{F}_{n,\rho(n)}(x) - f(x)\right| = \sum_{k=1}^{K} \sup_{x\in W_k} \left|\overline{F}_{n,\rho(n)}(x) - f(x)\right|. \tag{7.1}$$

We split each term

$$\begin{aligned}\sup_{x\in W_k} \left|\overline{F}_{n,\rho(n)}(x) - f(x)\right| = \sup_{x\in W_k} \left|\overline{F}_{n,\rho(n)}(x) - \overline{F}_{n,\rho(n)}(x_k)\right| + \sup_{x\in W_k} |f(x) - f(x_k)| \\ + \left|\overline{F}_{n,\rho(n)}(x_k) - f(x_k)\right|. \end{aligned} \tag{7.2}$$

Let $\varepsilon > 0$. The idea is to choose the radiuses r_k small enough to ensure that each term is controlled by a function of ε. Now, we make the idea precise. For all $k = 1, \dots, K$, the last term term can be made smaller that ε/K for n larger that some N_k using the pointwise convergence. For all $n \geq \max_{k\leq K} N_k$, and all $1 \leq k \leq K$, $\left|\overline{F}_{n,\rho(n)}(x_k) - f(x_k)\right| \leq \varepsilon/K$. The function f being continuous, it is uniformly continuous on every W_k. If we choose the W_k such that their radiuses are small enough (we may need to increase K), we can ensure that for all $1 \leq k \leq K$ $\sup_{x\in W_k} |f(x) - f(x_k)| \leq \varepsilon/K$. The first term on the r.h.s of (7.2) deserves more attention

$$\sup_{x\in W_k} \left|\overline{F}_{n,\rho(n)}(x) - \overline{F}_{n,\rho(n)}(x_k)\right| \leq \frac{1}{n}\sum_{i=1}^{n} \sup_{x\in W_k} \left|F_{i,\rho(n)}(x) - F_{i,\rho(n)}(x_k)\right|. \tag{7.3}$$

Now, for every $1 \leq k \leq K$, we want to apply Proposition 7.1.1 to the sequence of random variables $\left(\sup_{x\in W_k} |F_{n,m}(x) - F_{n,m}(x_k)|\right)_{n,m}$. Assumption ($\mathcal{H}$7.1) is clearly satisfied using Minkowski's inequality.

Let us define the sequence $(Y_{n,m})_{n,m}$ by

$$Y_{n,m} = \sup_{x\in W_k} |F_{n,m}(x) - F_{n,m}(x_k)| - \mathbb{E}\left[\sup_{x\in W_k} |F_{n,m}(x) - F_{n,m}(x_k)|\right],$$

satisfying $\mathbb{E}[Y_{n,m}] = 0$ and the assumptions of Proposition 7.1.1. Hence, it yields that

$$\lim_{n\to+\infty} \frac{1}{n}\sum_{i=1}^{n} \sup_{x\in W_k} \left|F_{i,\rho(n)}(x) - F_{i,\rho(n)}(x_k)\right| - \mathbb{E}\left[\sup_{x\in W_k} \left|F_{n,\rho(n)}(x) - F_{n,\rho(n)}(x_k)\right|\right] = 0. \tag{7.4}$$

From ($\mathcal{H}$7.4), we know that if the W_k are chosen small enough,

$$\sup_n \mathbb{E}\left[\sup_{x\in W_k} \left|F_{n,\rho(n)}(x) - F_{n,\rho(n)}(x_k)\right|\right] \leq \varepsilon/K. \tag{7.5}$$

Then, Combining with (7.3), (7.4) and (7.5) yields that $\sup_{x\in W_k} \left|\overline{F}_{n,\rho(n)}(x) - \overline{F}_{n,\rho(n)}(x_k)\right| \leq \varepsilon/K$. Going back to Equations (7.1) and (7.2), we deduce that for n large enough

$$\sup_{x\in W} \left|\bar{F}_{n,\rho(n)}(x) - f(x)\right| \leq 3\varepsilon,$$

which achieves the proof. ■

7.2 Stochastic approximation: the core martingale method

Consider a discrete time system satisfying for n large enough

$$X_{n+1} = X_n - \gamma_{n+1} u(X_n) - \gamma_{n+1} \delta M_{n+1}$$

where M_n is a martingale increment as studied in Chapter 2. We stick to the notation of that chapter.

Assume that the series $\sum_n \gamma_{n+1} \delta M_{n+1}$ converges .a.s., then we can introduce the auxiliary sequence $(X'_n)_n$ defined by

$$X'_n = X_n - \sum_{p \geq n+1} \gamma_p \delta M_p,$$

at least for n large enough.

Then, we obtain that $X'_{n+1} = X'_n - \gamma_{n+1} u(X_n)$. As $\lim_{n\to\infty} \sum_{p\geq n+1} \gamma_p \delta M_p = 0$ a.s., the pathwise behaviour of X_n and X'_n should be close for large enough n and smooth functions u, which means that this is reasonable to write

$$X'_{n+1} = X'_n - \gamma_{n+1} u(X'_n) + \gamma_{n+1}(u(X'_n) - u(X_n))$$

and think that the last term can be neglected. If one can prove, then one can apply the well–known ODE method introduced by Kushner and Clark [65] and further developed by Benaim [14], Benveniste et al. [16], Kushner and Yin [66] to carry out a purely pathwise analysis of the stochastic approximation.

To carry on our analysis, assume as in Chapter 2, that the function $x \longmapsto |x_\star|^2$ is a Lyapounov function. Then, we write from the definition of X'_n that

$$\left|X'_{n+1} - x_\star\right|^2 \leq \left|X'_n - x_\star\right|^2 - 2\gamma_{n+1}(X'_n - x_\star) \cdot u(X_n) + \gamma_{n+1}^2 \left|u(X_n)\right|^2 .$$

Now let n be large enough. We use the fact that X_n remains in a compact set and that u is locally bounded with local upper bound $\bar{u}$. Moreover $|X'_n - X_n| \leq \varepsilon$.

$$\left|X'_{n+1} - x_\star\right|^2 \leq \left|X'_n - x_\star\right|^2 - 2\gamma_{n+1}(X_n - x_\star) \cdot u(X_n) + \gamma_{n+1}^2 \bar{u}^2 + 2\gamma_{n+1}\varepsilon\bar{u}.$$

The rest of the analysis can be summarized as follows: if X'_n (or equivalently X_n) is far away from $x_\star$, the term $(X_n - x_\star) \cdot u(X_n) > 0$ drives X_n back to $x_\star$ until the remainder terms $\gamma_{n+1}^2 \bar{u}^2 + 2\gamma_{n+1}\varepsilon$ beat the retraction force. This cannot last long as the effect of the remainder terms vanish as soon as X_n starts to walk away from $x_\star$. As this reasoning can be made rigorous for any arbitrary small compact neighbourhood of $x_\star$, we conclude that X'_n converges to $x_\star$ and so does X_n.

Obviously, this methodology has a continuous time counterpart. Consider the process $dX_t = u(X_t)dt + dM_t$, where M is a martingale converging a.s., then the process X' solves $dX'_t = u(X_t)dt$, which does not have any martingale part and hence provided that X' and X are close enough, the behaviour of X is determined by the stability properties of u.

This methodology is successfully applied in discrete time in Chapter 2 and in continuous time in Chapter 6.

7.3 PNL: An open source numerical library

This library is a general purpose numerical library with a special emphasis towards high performance computing available under the LGPL (`https://pnlnum.github.io/pnl/`). I started developing PNL in 2007. I have been the main developer and welcomed a few external contributions since then. PNL was originally designed as the numerical library used by PREMIA (a financial pricer developed by INRIA, https://pnlnum.github.io/pnl/) but its scope quickly became much wider and I have kept introducing new features in connection to my research activities. The following publications strongly relied on the PNL library: [L-1], [L-2], [L-5], [L-8], [L-9], [L-11].

A wide range of routines are available on the following topics

- Complex Numbers;
- Cumulative Distribution Functions;
- Fast Fourier Transforms;
- Hyper Matrices;
- Laplace Inversion;
- Least-Squares fitting;
- Linear Algebra;
- List / array containers;
- MPI bindings to transparently pass the library objects on clusters;
- Multidimensional root Finding
- Multivariate polynomial regression;
- Numerical Integration;
- Optimization with inequality constraints including linear programming;
- Permutations;
- Random number generators (including a parallel Mersenne Twister);
- Approximation of special functions.

For some topics, state of the art libraries already existed. When their licenses were LGPL compatible, I integrated them into PNL. For instance, I can cite Lapack for linear algebra, QuadPack for numerical integration, MinPack for multi–dimensional root finding, Amos and Cephes for special function approximations.

Now, I would like to focus on two hot topics for future developments.

Multivariate regression Recently, I have mostly been working on multivariate regression, which is a key tool for all applications involving conditional expectations and backward induction. I have come up with a very efficient algorithm to build polynomial tensor representation. As an example, in Chapter 5, I managed to solve polynomial regressions up to degree 3 with 300 variates on clusters. The dimension of the vector space of polynomial tensors grows exponentially fast in particular when dealing with Wiener chaos expansions, for which the effective dimension is the number of time steps times the dimension of the underlying Brownian motion. Then, to tackle larger problems, it has become a burning issue to introduce adapted sparse polynomial representation. In the context of Wiener chaos expansion, some high order interactions between far time spaced Brownian increments should be irrelevant. Adding sparse and probably also local representations will be an important part of future developments. For this last example, the locality is obviously with respect to time and not space.

Parallel random number generators. Parallel random number generators are the cornerstone of any parallel stochastic algorithm. PNL already provides a parallel implementation of Mersenne

Twister based on [74]. Although this generator is very efficient on several hundreds of cores, it can not handle so efficiently architectures with several dozens of thousands of cores. To keep up with the evolution of massively parallel architectures and the growing sizes of problems, generators built on splitting approaches will be implemented, see [68] and [47].

These future developments will be extremely valuable for designing large scale parallel stochastic optimization methods, which will be of my main research topics for the coming years.

Chapter 8

Some prospects

As a conclusion, I would like to discuss some future works on a few topics presented in the document.

8.1 HPC for stochastic optimization

Many problems coming from financial mathematics eventually write as stochastic optimization problems. To tackle high dimensional problems, it has become a real stake to use high performance computing. This led us to advocate the use of sample average approximation rather than stochastic approximation in Chapter 5 to solve a minimization problem of the form

$$\inf_{\theta} \mathbb{E}[f(\theta, X)],$$

which is then replaced by its sample average approximation

$$\inf_{\theta} \sum_{i=1}^{n} f(\theta, X_i)$$

where the X_i's are i.i.d. with the distribution of X. When the approximated problem is strictly convex and twice differentiable, it can be solved very efficiently using a gradient descent approach with the descent direction given by the inverse of the Hessian matrix applied to the gradient, which ensures the decrease of the cost function. However, when the problem is only once differentiable or when the Hessian matrix is not tractable, computing an acceptable move becomes an issue. It turned out to be a major difficulty in the algorithm developed in Chapter 5. We also faced similar difficulties during the collaboration with *Mentor Graphics* on rare event simulation for electronic circuit design using importance sampling

In a deterministic context, line search techniques are commonly used to determine an acceptable move along a given direction. As line search requires several evaluations of the cost function, it becomes barely usable in a stochastic context when the dimension of the problem increases and adapting line search techniques to a probabilistic framework is an active research field, see for instance [73], which was recently published in the NIPS conference. A lot of computational time could be saved by not evaluating the complete cost function during the line search steps but only a rough approximation at least during the first iterations of the gradient method. Introducing such methods raises both theoretical questions on the convergence of the algorithm but also practical questions especially when the optimization problem is solved in a distributed environment. Collaborations with Franck Iutzeler and Jérôme Malick, who are specialists of deterministic optimization will be fruitful to derive new solutions making the most of both the stochastic and deterministic approaches.

8.2 A stochastic optimization point of view to BSDE

Solving backward stochastic differential equations is a challenging problem. Many techniques have been developed in the past 20 years based on at least one the following ingredients: Picard iterations, dynamic programming principle and the associated semi–linear PDE in the Markovian case. Recently, an approach based on Wiener chaos expansion has been developed.

Fruitful discussions with Philippe Briand led use to imagine a new approach to BSDE based on stochastic optimization. Consider the following BSDE

$$\bar{Y}_t = \xi + \int_t^T f(s, \bar{Y}_s, \bar{Z}_s)ds - \int_t^T \bar{Z}_s \cdot dB_s, \quad 0 \leq t \leq T \tag{8.1}$$

where B is a $d-$dimensional Brownian motion. If the pair $(\bar{Y}, \bar{Z})$ solves (8.1), then the unique solution Y of the standard SDE

$$Y_t = \bar{Y}_0 - \int_0^t f(s, Y_s, \bar{Z}_s)ds + \int_0^t \bar{Z}_s \cdot dB_s, \quad 0 \leq t \leq T, \tag{8.2}$$

satisfies $Y_T = \xi$ and $\bar{Y} = Y$.

This observation is the starting point of the new methodology, we want to develop, which consists in studying the BSDE (8.1) by solving the following minimization problem

$$\inf_{(y,Z)} \mathbb{E}\left[|Y_T - \xi|^2\right] \tag{8.3}$$

under the constraints

$$\begin{aligned} &Y_t = y - \int_0^t f(s, Y_s, Z_s)ds + \int_0^t Z_s \, dB_s, \quad 0 \leq t \leq T \\ &y \in \mathbb{R}^k, \quad Z \in \mathbb{L}^2([0,T] \times \Omega) \ (\mathcal{F})_{0 \leq t \leq T} \text{ adapted.} \end{aligned}$$

The pair $(\bar{Y}_0, \bar{Z})$ clearly solves the minimization problem. This new approach raises a wide range of questions going from the theoretical properties of the minimization problem to simulation issues. The numerical approach proposed here will directly benefit from the new advances of Section 8.1. The main advantage of this new approach is to rely only on forward simulations. Once the process Z is fixed, Y solves a forward SDE and can be discretized on a time grid using an Euler scheme for instance, provided that we can jointly simulate Z and B on the same time grid.

We believe that this approach is very promising and will enable a major breakthrough in the field of numerical simulation for high dimensional BSDE.

8.3 Dynamic programming principle and HPC

Many stochastic problems can be cast into a dynamic programming principle whose discretized version writes

$$\begin{cases} P_N^N = \psi(X_T) \\ P_n^N = \max(\psi(X_{t_n}), \mathbb{E}[P_{n+1}^N | \mathcal{F}_{t_n}]) \quad \text{for } n < N. \end{cases}$$

In principle, such problems can be solved exactly when the underlying process X is a discrete state space Markov chain even though it may become computationally intractable for high dimensional processes X. For one or two dimensional problems, this approach has been widely exploited to solve optimal stopping problems as it boils down to tree traversal. The exponential growth of the tree when

the size of X increases makes the use of such techniques far too computationally demanding in high dimensions.

Based on [91], we have already investigated the one dimensional case with Christophe Picard and we obtained promising results by traversing the tree in a way that isolates independent sub–trees (sub–problems), which can be solved in parallel. Extending this approach to multidimensional problems requires the development of new design patterns related to the geometric structure of the problem to come up with a generic parallel software for tree methods. The goal of all this research, which lies more on the computer science field, is to solve high dimensional backward stochastic differential equations using [21], in which the authors suggest to approximate the Brownian motion by a discrete random walk.

8.4 Stochastic modeling for ferro–magnets

The works on stochastic modeling for ferro–magnets we started with Stéphane Labbé, are only at their very beginning. We only studied the case of an isolated particle so far. When going from an isolated particle to a net of particles, the potential function governing the evolution of the system changes to take into account the interactions between particles, which makes the system far more complex as it admits several equilibrium positions. The first step is to analyse which kinds of limiting behaviours can be obtained from stochastic models with finitely many particles. In practice, particles tend to align by blocks, which are the equilibriums of the deterministic system. We will specifically focus on how to highlight switches between blocks, which corresponds to transitions from one equilibrium position to an other. This problem will be studied in a PhD thesis we will propose next year with S. Labbé.

The second step is to let the number of particles within a fixed volume increase to infinity. If properly scaled, we hope to obtain a stochastic partial differential equation governing the system, which can reproduce the thermal effects observed in practice. I have already had the opportunity to discuss some questions related to this extension with Andreas Prohl when he visited Laboratoire Jean Kuntzmann in April.

We ultimately aim at relating the variation of temperature observed in physical experiments to the quadratic variation of the stochastic model in order to understand how to control such a model, which is a burning issue in many applications such as wave protection, electronic compatibility, nano-electronics (see [1] for the deterministic case). The study of stochastic models for ferromagnetism raises both theoretical questions but also numerical questions as the simulation of such large systems is highly demanding especially when we focus on long term behaviours. A high performance C++ library will be developed to easily compare the different models and their behaviours in various contexts.

Bibliography

[1] S. Agarwal, G. Carbou, S. Labbé, and C. Prieur. Control of a network of magnetic ellipsoidal samples. *Mathematical Control and Related Fields*, 1(2):129–147, 2011. doi: 10.3934/mcrf.2011.1.129.

[2] L. Andersen and M. Broadie. Primal-dual simulation algorithm for pricing multidimensional american options. *Management Science*, 50(9):1222–1234, 2004.

[3] B. Arouna. Adaptative Monte Carlo method, a variance reduction technique. *Monte Carlo Methods Appl.*, 10(1):1–24, 2004.

[4] B. Arouna. Robbins-Monro algorithms and variance reduction in finance. *The Journal of Computational Finance*, 7(2), Winter 2003/2004.

[5] V. Bally and G. Pages. A quantization algorithm for solving multidimensional discrete-time optimal stopping problems. *Bernoulli*, 9(6):1003–1049, 2003.

[6] V. Bally and D. Talay. The law of the Euler scheme for stochastic differential equations. I. Convergence rate of the distribution function. *Probab. Theory Related Fields*, 104(1):43–60, 1996.

[7] O. Barndorff-Nielsen and N. Shephard. Modelling by lévy processess for financial econometrics. In O. Barndorff-Nielsen, S. Resnick, and T. Mikosch, editors, *Lévy Processes*, pages 283–318. Birkhäuser Boston, 2001.

[8] O. E. Barndorff-Nielsen and N. Shephard. Non-gaussian ornstein–uhlenbeck-based models and some of their uses in financial economics. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 63(2):167–241, 2001.

[9] O. E. Barndorff-Nielsen and R. Stelzer. The multivariate supou stochastic volatility model. *Mathematical Finance*, 23(2):275–296, 2013. doi: 10.1111/j.1467-9965.2011.00494.x.

[10] D. Belomestny. Solving optimal stopping problems via empirical dual optimization. *Ann. Appl. Probab.*, 23(5):1988–2019, 2013.

[11] D. Belomestny, C. Bender, and J. Schoenmakers. True upper bounds for Bermudan products via non-nested Monte Carlo. *Math. Finance*, 19(1):53–71, 2009.

[12] M. Ben Alaya and A. Kebaier. Central limit theorem for the multilevel monte carlo euler method. *Ann. Appl. Probab.*, 25(1):211–234, 02 2015.

[13] M. Ben Alaya, K. Hajji, and A. Kebaier. Importance sampling and statistical Romberg method. *Bernoulli*, 21(4):1947–1983, 2015.

[14] M. Benaim. A dynamical system approach to stochastic approximations. *SIAM Journal on Control and Optimization*, 34(2):437–472, 1996.

[15] M. Benaïm and M. W. Hirsch. Stochastic approximation algorithms with constant step size whose average is cooperative. *Annals of Applied Probability*, 9(1):216–241, 1999.

[16] A. Benveniste, M. Métivier, and P. Priouret. *Stochastic approximations and adaptive algorithms*. Springer-Verlag, 1990.

[17] D. P. Bertsekas. Stochastic optimization problems with nondifferentiable cost functionals. *J. Optimization Theory Appl.*, 12:218–231, 1973.

[18] P. Billingsley. *Probability and measure*. John Wiley & Sons, 2008.

[19] N. Bouleau and D. Lépingle. *Numerical methods for stochastic processes*. Wiley Series in Probability and Mathematical Statistics: Applied Probability and Statistics. John Wiley & Sons Inc., New York, 1994. A Wiley-Interscience Publication.

[20] S. Boyd, L. Xiao, and A. Mutapcic. Subgradient methods. *lecture notes of EE392o, Stanford University, Autumn Quarter*, 2004:2004–2005, 2003.

[21] P. Briand, B. Delyon, and J. MÈmin. Donsker-type theorem for BSDEs. *Electron. Comm. Probab*, 6:1–14, 2001.

[22] D. R. Brillinger. A particle migrating randomly on a sphere. *Journal of Theoretical Probability*, 10(2):429–443, 1997.

[23] M. Broadie and P. Glasserman. A stochastic mesh method for pricing high-dimensional american options. *Journal of Computational Finance*, 7:35–72, 2004.

[24] W.-F. Brown. *Magnetostatic Principles in Ferromagnetism*. North-Holland, 1962.

[25] W. F. Brown. Thermal fluctuations of a single-domain particle. *Phys. Rev.*, 130:1677–1686, Jun 1963. doi: 10.1103/PhysRev.130.1677.

[26] J. F. Carriere. Valuation of the early-exercise price for options using simulations and nonparametric regression. *Insurance: mathematics and Economics*, 19(1):19–30, 1996.

[27] A. Carverhill and N. Webber. American options: theory and numerical analysis. *Options: recent advances in theory and practice*, pages 80–94, 1990.

[28] H. F. Chen. *Stochastic approximation and its applications*, volume 64 of *Nonconvex Optimization and its Applications*. Kluwer Academic Publishers, Dordrecht, 2002.

[29] H. F. Chen and Y. M. Zhu. *Stochastic Approximation Procedure with randomly varying truncations*. Scientia Sinica Series, 1986.

[30] H. F. Chen, G. Lei, and A. J. Gao. Convergence and robustness of the Robbins-Monro algorithm truncated at randomly varying bounds. *Stochastic Process. Appl.*, 27(2):217–231, 1988.

[31] M. H. A. Davis and I. Karatzas. A deterministic approach to optimal stopping. In *Probability, statistics and optimisation*, Wiley Ser. Probab. Math. Statist. Probab. Math. Statist., pages 455–466. Wiley, Chichester, 1994.

[32] P.-T. de Boer, D. P. Kroese, S. Mannor, and R. Y. Rubinstein. A tutorial on the cross-entropy method. *Ann. Oper. Res.*, 134:19–67, 2005.

[33] B. Delyon. General results on the convergence of stochastic algorithms. *IEEE Transactions on Automatic Control*, 41(9):1245–1255, 1996.

[34] B. Delyon, M. Lavielle, and E. Moulines. Convergence of a stochastic approximation version of the EM algorithm. *Ann. Statist.*, 27(1):94–128, 1999.

[35] S. Dereich. Multilevel Monte Carlo algorithms for Lévy-driven SDEs with Gaussian correction. *Ann. Appl. Probab.*, 21(1):283–311, 2011.

[36] D. Duffie and P. Glynn. Efficient Monte Carlo simulation of security prices. *Ann. Appl. Probab.*, 5(4):897–905, 1995.

[37] M. Duflo. *Random Iterative Models*. Springer-Verlag Berlin and New York, 1997.

[38] V. Dung Doan, A. Gaiwad, M. Bossy, F. Baude, and I. Stokes-Rees. Parallel pricing algorithms for multimensional bermudan/american options using Monte Carlo methods. *Mathematics and Computers in Simulation*, 81(3):568–577, 2010.

[39] J.-C. Fort and G. Pages. Asymptotic behavior of a markovian stochastic algorithm with constant step. *SIAM journal on control and optimization*, 37(5):1456–1482, 1999.

[40] C. Geiss and C. Labart. Simulation of BSDEs with jumps by wiener chaos expansion. *Stochastic Processes and their Applications*, 2016. URL `http://dx.doi.org/10.1016/j.spa.2016.01.006`.

[41] M. B. Giles. Improved multilevel Monte Carlo convergence using the Milstein scheme. In *Monte Carlo and quasi-Monte Carlo methods 2006*, pages 343–358. Springer, Berlin, 2008.

[42] M. B. Giles and L. Szpruch. Antithetic multilevel Monte Carlo estimation for multi-dimensional SDEs without Lévy area simulation. *Ann. Appl. Probab.*, 24(4):1585–1620, 2014.

[43] M. B. Giles, D. J. Higham, and X. Mao. Analysing multi-level Monte Carlo for options with non-globally Lipschitz payoff. *Finance Stoch.*, 13(3):403–413, 2009.

[44] P. Glasserman. *Monte Carlo methods in financial engineering*, volume 53 of *Applications of Mathematics (New York)*. Springer-Verlag, New York, 2004. Stochastic Modelling and Applied Probability.

[45] K. Hajji. *Accélération de la méthode de Monte Carlo pour des processus de diffusions et applications en Finance*. PhD thesis, Université Paris 13, 2014.

[46] P. Hall and C. Heyde. *Martingale Limit Theory and its Application*. Academic Press, 1980.

[47] H. Haramoto, M. Matsumoto, T. Nishimura, F. Panneton, and P. L'Ecuyer. Efficient jump ahead for f2-linear random number generators. *INFORMS J. on Computing*, 20:385–390, July 2008.

[48] M. B. Haugh and L. Kogan. Pricing american options: a duality approach. *Operations Research*, 52(2):258–270, 2004.

[49] S. Heinrich. Monte Carlo complexity of global solution of integral equations. *J. Complexity*, 14 (2):151–175, 1998.

[50] S. Heinrich. Multilevel monte carlo methods. *Lecture Notes in Computer Science, Springer-Verlag*, 2179(1):58–67, 2001.

[51] S. Heinrich and E. Sindambiwe. Monte carlo complexity of parametric integration. *J. Complexity*, 15(3):317–341, 1999. Dagstuhl Seminar on Algorithms and Complexity for Continuous Problems (1998).

[52] J. Jacod and P. Protter. Asymptotic error distributions for the Euler method for stochastic differential equations. *Annals of Probability*, 26(1):267–307, 1998.

[53] J. Jacod and A. Shiryaev. *Limit Theorems for Stochastic Processes*. Springer-Verlag Berlin, 1987.

[54] F. Jamshidian. The duality of optimal exercise and domineering claims: a Doob-Meyer decomposition approach to the Snell envelope. *Stochastics*, 79(1-2):27–60, 2007.

[55] B. Jourdain. *Advanced Financial Modelling*, chapter Adaptive variance reduction techniques in finance, pages 205–222. Radon Series Comp. Appl. Math 8. Walter de Gruyter, 2009.

[56] R. Kawai. Adaptive Monte Carlo variance reduction with two-time-scale stochastic approximation. *Monte Carlo Methods Appl.*, 13(3):197–217, 2007.

[57] R. Kawai. Adaptive Monte Carlo variance reduction for Lévy processes with two-time-scale stochastic approximation. *Methodol. Comput. Appl. Probab.*, 10(2):199–223, 2008.

[58] R. Kawai. Optimal importance sampling parameter search for Lévy processes via stochastic approximation. *SIAM J. Numer. Anal.*, 47(1):293–307, 2008.

[59] A. Kebaier. Statistical Romberg extrapolation: a new variance reduction method and applications to option pricing. *Ann. Appl. Probab.*, 15(4):2681–2705, 2005.

[60] R. Khasminskii. *Stochastic stability of differential equations*, volume 66. Springer Science & Business Media, 2011.

[61] S. Kim and S. G. Henderson. Adaptive control variates. In *Proceedings of the 2004 Winter Simulation Conference*, 2004.

[62] S. Kim and S. G. Henderson. Adaptive control variates for finite-horizon simulation. *Math. Oper. Res.*, 32(3):508–527, 2007.

[63] A. Kolodko and J. Schoenmakers. Upper bounds for bermudan style derivatives. *Monte Carlo Methods and Applications mcma*, 10(3-4):331–343, 2004.

[64] S. G. Kou. A jump-diffusion model for option pricing. *Manage. Sci.*, 48(8):1086–1101, 2002.

[65] H. J. Kushner and D. S. Clark. *Stochastic approximation methods for constrained and unconstrained systems*, volume 26 of *Applied Mathematical Sciences*. Springer-Verlag, New York, 1978.

[66] H. J. Kushner and G. G. Yin. *Stochastic approximation and recursive algorithms and applications*, volume 35 of *Applications of Mathematics (New York)*. Springer-Verlag, New York, second edition, 2003. Stochastic Modelling and Applied Probability.

[67] Lamberton. Brownian optimal stopping and random walks. *Applied Mathematics & Optimization*, 45(3):283–324, 2002.

[68] P. L'Ecuyer and S. Côté. Implementing a random number package with splitting facilities. *ACM Trans. Math. Softw.*, 17(1):98–111, 1991.

[69] M. Ledoux and M. Talagrand. *Probability in Banach spaces*, volume 23 of *Ergebnisse der Mathematik und ihrer Grenzgebiete (3) [Results in Mathematics and Related Areas (3)]*. Springer-Verlag, Berlin, 1991. Isoperimetry and processes.

[70] V. Lemaire and G. Pagès. Unconstrained Recursive Importance Sampling. *Ann. Appl. Probab.*, 20(3):1029–1067, 2010.

[71] D. Lépingle. Sur le comportement asymptotique des martingales locales. *Séminaire de probabilités de Strasbourg*, 12:148–161, 1978.

[72] F. Longstaff and R. Schwartz. Valuing American options by simulation : A simple least-square approach. *Review of Financial Studies*, 14:113–147, 2001.

[73] M. Mahsereci and P. Hennig. Probabilistic line searches for stochastic optimization. In *Proceedings of the 28th International Conference on Neural Information Processing Systems*, NIPS'15, pages 181–189, Cambridge, MA, USA, 2015. MIT Press.

[74] M. Matsumoto and T. Nishimura. Dynamic creation of pseudorandom number generators. In *Monte Carlo and quasi-Monte Carlo methods 1998 (Claremont, CA)*, pages 56–69. Springer, Berlin, 2000.

[75] D. Nualart. Analysis on Wiener space and anticipating stochastic calculus. In B. Springer-Verlag, editor, *Lectures on Probability Theory and Statistics (Saint- Flour, 1995)*, pages 123–227. 1998.

[76] B. T. Polyak. Introduction to optimization. *Optimization Software*, 1987.

[77] R. Rebolledo. Central limit theorems for local martingales. *Z. Wahrsch. Verw. Gebiete*, 51(3): 269–286, 1980.

[78] D. Revuz. *Probabilités*. Hermann, 1997.

[79] H. Robbins and S. Monro. A stochastic approximation method. *Ann. Math. Statistics*, 22:400–407, 1951.

[80] L. C. G. Rogers. Monte Carlo valuation of American options. *Math. Finance*, 12(3):271–286, 2002.

[81] L. C. G. Rogers. Dual valuation and hedging of Bermudan options. *SIAM J. Financial Math.*, 1: 604–608, 2010.

[82] L. C. G. Rogers and D. Williams. *Diffusions, Markov processes, and martingales. Vol. 2*. Cambridge Mathematical Library. Cambridge University Press, Cambridge, 2000.

[83] R. Y. Rubinstein and A. Shapiro. *Discrete event systems*. Wiley Series in Probability and Mathematical Statistics: Probability and Mathematical Statistics. John Wiley & Sons Ltd., Chichester, 1993. Sensitivity analysis and stochastic optimization by the score function method.

[84] J. Schoenmakers. *Robust Libor modelling and pricing of derivative products*. CRC Press, 2005.

[85] J. Schoenmakers, J. Zhang, and J. Huang. Optimal dual martingales, their analysis, and application to new algorithms for bermudan products. *SIAM Journal on Financial Mathematics*, 4(1): 86–116, 2013.

[86] Y. Su and M. C. Fu. Optimal importance sampling in securities pricing. *Journal of Computational Finance*, 5(4):27–50, 2002.

[87] D. Talay and L. Tubaro. Expansion of the global error for numerical schemes solving stochastic differential equations. *Stochastic Anal. Appl.*, 8(4):483–509 (1991), 1990.

[88] J. A. Tilley. Valuing american options in a path simulation model. *Transactions of the Society of Actuaries*, 45(83):104, 1993.

[89] J. Tsitsiklis and B. V. Roy. Regression methods for pricing complex American-style options. *IEEE Trans. Neural Netw.*, 12(4):694–703, 2001.

[90] W. Whitt. Proofs of the martingale FCLT. *Probab. Surv.*, 4:268–302, 2007.

[91] N. Zhang, A. Roux, and T. Zastawniak. Parallel binomial valuation of american options with proportional transaction costs. In *Advanced Parallel Processing Technologies*, volume 6965 of *Lecture Notes in Computer Science*. Springer Berlin Heidelberg, 2011.